Stanford eCorner

New Opportunities in Search Engine Technology

06-11-2008

URL: https://stvp.stanford.edu/blog/videos/new-opportunities-in-search-engine-technology

When Alta Vista launched, it indexed about 50 percent of the Web. Years later, Google launched at the same capacity. But the opportunity today in search engine technology, says Cuil co-founder Anna Patterson, is that the outlay for the hardware required to search the ever-growing pool of content has become prohibitively expensive. In this clip, Patterson describes Cuil's alternative, mathematical algorithm to "scatter gather" search that is both more thorough and more efficient, allowing just 140 machines to log 124 billion pages.

Transcript

The Web has grown super exponentially because it still looks like an exponential even plotted on log scale and you can see when each of these properties came on the market.. So when AltaVista came on the market, it indexed over 50% of the Web.. And when Google launched, it indexed about 50% of the Web.. So as the Web has grown, people need to buy a certain amount of hardware in order to index a certain amount of information and so obviously you can't just invest exponentially in hardware.. So search engines really haven't kept up with exponential growth of the Web.. So we thought that provided an opportunity for Cuil, because rather than buying more machines, Tom, our co-founder actually did some mathematics in order to come up with a mathematical model about how to build a search engine differently.. So standard search engine is built as a scattered-gather architecture, I don't know how many people know that terminology, but basically it would be like asking every single one of you the same question and you guys emailing me back the answer.. So you scatter the query and you gather up the answer.. So he came up with a representation so that you didn't have to scatter the query anywhere; that over 95% of queries could just go to one machine.. And with that, it means that you can have architecture that's entirely on disc instead of in memory..

If I have to ask every single one of you a query and you're all computers, then I'm going to have to use the memory because it's going to take too long in the latency of the network to get the question out to you, how do you compute the answer, and get it back to me.. So the latency is high so the computation on each of the machine needs to be low.. So he said, well, let's spend the latency instead going to disc and doing lots of computation.. So it does lots of computation on the mathematical model in order to figure out the answer.. So that's why on 140 machines, we can serve 125 billion pages.. And trust me, Yahoo! and Google use more...