Carnegie Mellon algorithm identifies top 100 blogs for news

November 19 2007

Being among the first to pick up on Internet news and gossip and rapidly detecting contamination anywhere in a water supply system are similar problems, at least from a computer scientist's point of view. Both can be solved with a versatile algorithm developed by Carnegie Mellon University researchers.

Using a problem-solving method called the Cascades algorithm, Carlos Guestrin, assistant professor of computer science and machine learning, and his students compiled a list of the best 100 blogs to read to find the biggest news on the Web as early as possible, www.blogcascades.org/ . It includes well-known blogs, such as Instapundit and Boing Boing, but also some more obscure ones like Watcher of Weasels and Don Surber.

"The goal of our system when looking at blogs is to detect the big stories as early on and as close to the source as possible," Guestrin said. He, Andreas Krause and Jure Leskovec, doctoral students in computer science and machine learning, respectively, analyzed 45,000 blogs (those that actively link to other blogs) to compile the list, checking the time stamps to determine where news items were being posted first.

But reading even 100 blogs, many of them with numerous postings, may be more than many Web surfers can handle. Recasting the problem, the researchers used their algorithm to compile a list of blogs if a person wanted to read only 5,000 postings. This list is quite different, with "summarizer" blogs, such as The Modulator and Anglican predominating.

Similarly, Guestrin and his students used the same algorithm to determine the optimal number and placement of sensors for detecting the introduction and spread of contaminants in a municipal water supply. Their report on the blog and water system case studies, “Cost-Effective Outbreak Detection in Networks,” was presented at the Association for Computing Machinery’s International Conference on Knowledge Discovery and Data Mining earlier this year.

“Nothing demonstrates the versatility of Carlos’ algorithm better than its ability to solve these two difficult and seemingly different problems,” said Randal E. Bryant, dean of Carnegie Mellon’s School of Computer Science. “It’s a credit to Carlos’ insight and inventiveness, but also a testament to the power of computational thinking. Computer scientists increasingly are developing common methods for solving problems that apply across any number of disciplines.”

Guestrin began work on the Cascades algorithm in 2004 to find a way to balance the cost of collecting information with the need for collecting the information early and close to its source. Initially, this addressed problems in designing wireless sensor networks — a technology that potentially can monitor such important conditions as water quality, building temperature, vital signs of nursing home residents, algal blooms in lakes and the structural integrity of bridges. In all of these cases, deploying the wrong number of sensors or putting them in the wrong places wastes money and produces poor information.

The algorithm allows for near-optimal placement of sensors by exploiting a property called submodularity. Simply put, submodularity means there is a diminishing return associated with adding sensors — adding a sensor to a five-sensor network has much more impact than adding a sensor to a 10,000-sensor network. The algorithm also takes into account the property of locality — the idea that sensors that are far apart provide almost independent information.

Work by Guestrin and his group is now focusing on detecting pollution in lakes and rivers and ensuring performance quality on citywide Wi-Fi networks. “This project represents a nice blend of theoretical understanding and a lot of engineering effort to make the whole thing work,” he said. “It’s a nice theory applied to larger, real-world data. It’s cross fertilization and interdisciplinary thinking in the true Carnegie Mellon tradition.”

Work on developing the Cascade algorithm has been supported by the National Science Foundation, Intel, Microsoft, the Sloan Foundation, PITA, IBM and Hewlett-Packard.

Source: Carnegie Mellon University