
September 03, 2008

Semantic Supercomputing Reaps Competitive Advantage from Patent Data

by David Rich, Vice President of Marketing, Interactive Supercomputing Inc.

Mark Twain wrote, "A country without a patent office and good patent laws is just a crab, and couldn't travel any way except sideways and backwards." But a good patent infrastructure without good ways to search it does not move a country's industry forward.

Twain should know. He fought a protracted patent dispute with another man in 1871 over the invention of an elastic vest strap. Twain ultimately prevailed, but could have saved himself a great deal of money, time and frustration had he known about the competing patent beforehand.

Every year billions of dollars are wasted on research and development of inventions that are already protected by patent law -- an estimated \$20 billion in the U.S. and &euro60 billion in Europe, which equals roughly the combined annual revenues of Microsoft and Apple. In fact, these computing titans themselves have fought costly intellectual property wars due to poor patent intelligence, such as the 2004 patent dispute over the iPod user interface, which Apple ultimately lost to Microsoft.

It's no wonder patent information is so costly and difficult to divine. The volume of patent data is overwhelming. The world's collection of patents comprises the largest information repository of the most important achievements of humanity. Since the first patent was issued for a Venetian statue in 1471, 60 million patents have been awarded around the world, with four million patents actively in force today worldwide. And 800,000 new inventions are registered every year. While the data is public, current search tools are inconvenient and inadequate to the needs of professionals. And even if you solve the patent retrieval problem, it's not enough: researchers today need integrated views of correlated patent information, such as corporate affiliations, scientific information, prior art documents, and breaking news on intellectual property.

To address this challenge, researchers are developing computationally-intensive natural language processing (NLP) algorithms in the new field of semantic supercomputing. One company that is tapping the new technology is Vienna, Austria-based Matrixware Information Services (www.matrixware.com). The firm is combining HPC systems with Interactive Supercomputing, Inc.'s (ISC's) Star-P software to tackle the ever-growing challenge of finding patent information hidden in the world's vast patent databases and libraries.

Patents and intellectual property play an increasingly important role as intangible assets of industrial corporations. Over 250,000 companies worldwide depend on patent data. Consequently, professional management of patents and precise retrieval of patent information are essential business processes for industries around the globe.

Companies pioneering semantic computing typically employ teams of computer engineers, mathematicians, linguists and patent specialists to help companies mine patent repositories for intellectual property information.

The semantic supercomputing techniques and HPC technology they utilize enable the users to retrieve relevant patent information faster, more easily and at less cost.

Matrixware, for example, employs multicore SGI Altix 4700 blade servers and Linux clusters running Star-P to develop and run its NLP algorithms on terabyte-scale patent data sets. Star-P enables Matrixware's team to continuously code and refine NLP algorithms on their desktops using Python or MATLAB, and then run them interactively on HPC systems with little to no modification. The semantic supercomputing model eliminates the need to re-program applications in C, Fortran or MPI in order to run on parallel systems, resulting in huge productivity gains.

Patent retrieval presents two levels of computational challenges. The first challenge is data centric. The patent information is dispersed among several hundred repositories, dating back as far as the 1700s. These diverse patent collections have evolved through 200 generations of methods of storing documents between then and today. Some of the information is digital data; other is derived from documents that have been scanned and converted with OCR systems, and others are just plain document images. Researchers must wrestle with enormous gaps and inconsistencies in the format of 100 million documents.

Another challenge is database centric. Today, most patent data is stored in relational databases. But the art of managing patent information is based on 4,000 years of library science methods, which conflict with the restrictions imposed by relational databases. This severely limits the accessibility to the data.

For example, most patent documents are classified by a taxonomy scheme set up by the World Intellectual Property Organization (WIPO) that contains approximately 70,000 classes, called the Intellectual Property and Technology Commercialization (IPTC) taxonomy. It ranges widely from chemical to mechanical patent classifications with many sub-classifications beneath each major class (e.g. automotive being a sub-class of mechanical).

"We wanted to see if there are specific terms that are characteristic for specific classes within the taxonomy," said Matrixware CEO Francisco Weber. His team tried using relational databases running on a conventional server, taking a sampling of about 1.5 million patent documents, from which they extracted 10 billion terms. They then created a simple database join to aggregate the terms according to their classifications. The result was a database join of 1.5 million X 10 billion rows. "We ended up busting every commercial database system we tried," he added.

To solve the problem, Matrixware developed the Alexandria System -- a central storage repository for the raw data as well as for enriched data running on the HPC systems. It takes a different approach to storing and managing large amounts of document data. The data access of Alexandria is modeled along the well-established library science methods and embedded into a workflow system. The Alexandria server also provides the user with exact and constantly updated document counts in the collections from which the researcher retrieves.

To make patent information usable, accessible and meaningful, the Alexandria System recursively generates metadata from data as well as metadata from metadata. These refinement processes continuously feed and update the Alexandria repository and allow users to actively "cultivate the corpus," to use the industry expression for creating a rich collection of linguistic data.

To provide a front-end development framework to Alexandria, Matrixware created a software infrastructure called the Leonardo Ecosystem. Within this framework, technologists can simultaneously create and refine new search tools and methods, as well as collaborate with other users in the user community to solve problems. This benefits users by allowing them to choose the best available tool for specific information needs and existing workflows.

The huge memory models required by the patent corpus contained in Alexandria required an 80-processor node

SGI Altix system with 380 GB of memory. The Alexandria system presented two computational challenges that could only be addressed by HPC systems. The first was at the pure textstring level. The process of splitting text, extracting and tokenizing words from a collection of 1.5 million patent documents generates 10 billion terms, requiring memory models of multiple terabytes for processing. Moreover, the research process is inherently iterative and experimental, requiring constant refinement of the NLP algorithms and repetitive batch runs on the computer. Consequently, supercomputing speeds were necessary to make the work flow reasonable.

The second challenge was handling the huge matrices. The text strings that are extracted create counts of occurrences of terms within the patent documents, which are encoded numerically. This results in matrices of up to 10 million by 10 billion. Applying the algorithms to these matrices required a software platform that could scale to whatever extent the data required.

Semantic supercomputing processes patent data by its contextual meaning to turn it into valuable information for users. Its purpose is to boost their productivity and open up new opportunities for them using intellectual property information. But while users are typically experts in information retrieval, they are not parallel programming experts. Semantic supercomputing enables them to tap the power of HPC systems to refine and run their natural language processing applications as well as to improve the data quality of patent repositories.

About the Author

David Rich is the Vice President of Marketing at Interactive Supercomputing. David brings to ISC more than 23 years of marketing, sales and support experience in both large and entrepreneurial high tech companies. At AMD he directed the company's entry into the HPC cluster market and secured large wins such as the Red Storm system at Sandia National Laboratories and the Dawning 4000A at the Shanghai Supercomputer Center. While at AMD, he served as president of the HyperTransport Consortium, a standards organization for high-speed interconnect technology. David's earlier experience includes being the founding manager of the TotalView product line, which has become the de facto standard for parallel and distributed debugging. He served as vice president of Fujitsu System Technologies, which developed high-speed networking technology that was a pre-cursor to InfiniBand. His parallel processing experience started at BBN Technologies where he worked on the Butterfly series of computers. David received a bachelor's degree in computer science from Brown University.
