

On Scalable (Computer-Based) Information Systems

by Ophir Frieder

Editor's Note: The ASIS&T Research Award, one of the society's most prestigious honors, recognizes an individual's outstanding contributions to information science research. In recent years the *Bulletin* has asked each honoree to write an article for us discussing his or her work and its significance. Dr. Ophir Frieder, the IITRI Chair Professor of Computer Science and the director of the Information Retrieval Laboratory at the Illinois Institute of Technology, is the 2007 recipient of the Research Award.

Information systems permeate every aspect of daily life. From our personalized phone directories to our bank accounts, from our medical records to our online restaurant searches, our reliance on information systems is simply a fact of life. So what exactly are information systems? According to an October 1, 2007, Wikipedia search, an *information system* is “the system of persons, data records and activities that process the data and information in a given organization, including manual processes or automated processes. Usually the term is used erroneously as synonymous for computer-based information systems, which is only the information technologies component of an information system.” As noted, the terms *information systems* and *computer-based information systems* are often used interchangeably, even if erroneously, and both do often refer only to the information technology components.

Given the above terminology, what constitutes *scalable* computer-based information systems? The definition for *scalability*, according to *Wiktionary*, is “the ability to support the required quality of service as the system load increases without changing the system.” Thus, scalable computer-based information systems are those systems that continue to sustain the required quality of service even under an increased load. This definition, however,

Ophir Frieder is IITRI Chair Professor of Computer Science and director of the Information Retrieval Laboratory at the Illinois Institute of Technology, Department of Computer Science, 10 W. 31st St., Chicago, IL 60616; phone: 312-567-5143; email: ophir<at>ir.iit.edu

begs the questions: “What load must a computer-based information system efficiently process so as to merit the distinction of scalable?” and “What is the required quality of service?” Unfortunately, the answers to these questions are rather vague and are complicated by the inherent natures of scalability and quality.

Examining first scalability, the notion of *size*, and hence, *increased system load*, varies according to time (and hence technology) and application domain. In the 1960s, a test collection of roughly 1400 documents and 225 queries known as the Cranfield collection (“The Cranfield Tests on Index Language Devices,” reprinted in *Readings in Information Retrieval* by Sparck Jones & Willett, 1997) was considered the state of the art. In 1992, the Text REtrieval Conference (TREC) first introduced a large text collection benchmark. This collection was approximately two GB in size; one GB was used for training, the other for evaluation. Many if not most of the participants found the volume of data to be excessively large. In 1997, TREC introduced a track known as the “Very Large Corpus,” whose collection was comprised of just slightly more than 20 GB, and in 2004 a newer very large corpus track was introduced, called the “Terabyte Track.” Terabyte Track is a bit of a misnomer, however, since the test data consists of a 426 GB collection of roughly 25 million documents taken from the [dot]gov World Wide Web domain. As can be seen by the steady and significant growth in these numbers, the implication of size, and thereby increased load, has evolved over time.

Another aspect of scalability is the domain. In the data-warehousing domain, systems processing multiple hundreds of terabytes are commonplace. Petabyte collections are often discussed. In the text domain, however, according to worldwidewebsize.com on October 1, 2007, the size of the indexed World Wide Web is just slightly shy of 23 billion pages. This estimate is roughly a median size estimate, as common WWW size estimates range in the 15 to 30 billion indexed pages. Given that the leading search engines each index less than half of these pages (Google is estimated to index less than 9 billion pages), clearly, even the Web search engines store and index less than a hundred terabytes of text. Note, however, that search engines are prime examples of systems that process large volumes of text – especially when we consider that the print collections of the U.S. Library of Congress are often estimated at only 10 terabytes. In the highly complex scanned-document domain, searching even single digit terabyte sized collections poses a research nightmare let alone an industrial day-to-day practical reality. Thus, the composition of the application domain and technology available impacts what constitutes scalable.

Addressing the issue of quality of services is even more nebulous. Users always want more. They want more intuitive interfaces to more accurate systems that return the answers even faster. User satisfaction can be improved but never fulfilled. What is clear, however, is that to meet this required quality of service attention must focus not only on the system but also on the users themselves.

Having loosely, or at least intuitively, outlined what scalable computer-based information systems are, where do I fit within this domain? The answer to this question, likewise, has evolved over time. Given the ASIS&T community focus, I will limit my discussion to only the traditional information retrieval domain and partition this exploration into efficiency, accuracy and user-focused efforts. I shall proceed chronologically, and hence, start with efficiency.

With a computer architecture and parallel processing background, initially I mistakenly believed that the “way to scalability” in information retrieval was via computer hardware technology. Thus, as reported in a co-authored 1991 article, “Exploiting Parallelism in Pattern Matching: An Information

Retrieval Application”, in the *ACM Transactions on Information Systems (ACM TOIS)*, we developed a VLSI (very large systems integration)-based approach for what was then considered high speed text filtering. The approach relied on a custom-designed computer chip that supported multiple string comparators. All the comparators were identical, operated in parallel and implemented a novel early mismatch detection feature that compared the query terms against the text to identify relevant documents. However, the required custom-specific implementation rendered the approach obsolete before any wide practical acceptance occurred.

Having realized (the hard way) the inadequacy of customized-systems-based solutions, I focused on the use of conventional scalable architectures, namely, parallel processing engines. Two efforts ensued, both of which improved the efficiency, and in some cases also the accuracy, of clustering algorithms. In a 1997 co-authored *JASIS* article entitled, “Clustering and Classification of Large Document Bases in a Parallel Environment,” we designed, implemented and evaluated a parallel single-pass clustering approach that sustained nearly theoretically optimal processing times. More recently, in a co-authored 2007 *JASIST* article entitled, “Exploiting Parallelism to Support Scalable Hierarchical Clustering,” we propose and evaluate nearly equally efficient parallel hierarchical and Buckshot clustering approaches. The parallel clustering solutions partition the data, and hence the workload, across multiple processing sites (or nodes), all working in unison. Their approach mimics the computer science rendition of “two heads are better than one!” Our single-pass parallel clustering approach was adopted into daily practice by one of our sponsors.

Another approach to support scalability is to reduce (filter) the data early on so that only a portion of them is involved in successive computations. In a co-authored 2002 *ACM TOIS* article entitled, “Collection Statistics for Fast Duplicate Document Detection,” we specify an approach based on collection statistics that detects near duplicate documents. Terms are selectively chosen to represent each document. An encoding that is based on a mathematical function of a sorted ordering of these retained terms represents each document. All document encodings are efficiently compared and similar documents are detected. By eliminating similar documents, the number of documents

processed in successive stages is reduced, and hence, so is the workload. Approaches based on this effort are commercially deployed for spam detection and removal.

Of no less and arguably more criticality than efficiency is accuracy. To that end, I developed approaches that integrated different data types as well as multiple search engines. In co-authored 1997 and 1999 *JASIS* articles entitled, “Integrating Structured Data and Text: A Relational Approach” and “A Parallel Relational Database Management System Approach to Relevance Feedback in Information Retrieval,” we describe an approach that relies on strictly conventional relational database technology to support full information retrieval. That is, the main retrieval models, namely, the Boolean, vector space and probabilistic models, are all supported. Passage based, n-gram based and proximity-based retrieval and relevance feedback implementations are likewise supported. Furthermore, as the approach relies on strictly relational database scripts expressed in standard SQL, parallel database systems efficiently support the integrated querying of relational data and text, and any advancement in parallel database systems directly benefits information retrieval systems. Multiple vendors use this approach commercially.

Again improving effectiveness, in a co-authored 2004 *JASIS* article entitled, “On Fusion of Effective Retrieval Strategies in the Same Information Retrieval System,” we determined the effects on accuracy of varying individual processing components within an integrated multiple independent search engine system. We also used these experiments to characterize situations where fusion of disparate processing components was likely to improve system accuracy, thereby allowing a targeted fusion approach that resulted in more efficient retrieval systems. Both integration efforts supported the processing of larger collections either via reliance on technology well known to support large data volumes (relational database management systems) or merging the results of a partitioned, and hence more manageable, set of searches.

The final efforts I describe focus on the users, in terms of trying to better understand, predict and interact with them. Over time, my understanding of user needs has grown, but is still incomplete. That said, some of my efforts occurred haphazardly over time and are not necessarily presented chronologically.

To better understand user needs, as presented in a co-authored 2007 *JASIS* article entitled, “Temporal Analysis of a Very Large Topically Categorized Web Query Log,” we analyzed the usage patterns of many millions of queries of a commercial search engine. Some of the usage patterns were predictable; however, some were not and were subsequently used to improve retrieval.

In a co-authored 2007 *ACM TOIS* article entitled, “Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs,” we predict user intent and improve accuracy. Using the approach described, we sustain comparable accuracy using automated means on short web queries without any additional context as compared to more computationally intensive techniques that require and rely on additional context information. The ability to automatically predict, and hence interpret, user meaning simplifies the user’s query specification task.

To better interact with the user, as described in a co-authored 1999 *JASIS* article entitled, “SENTINEL: A Multiple Engine Information Retrieval and Visualization System,” we developed SENTINEL. SENTINEL was configured with multiple fused search engines that interacted with the user via a three-dimensional, stereophonic visualization component. As a computer scientist, I had always focused on the computer system. This effort, although described last, was my wake-up call of the importance of the “user in the loop,” and hence, finally guided me to an appreciation of the difference between the world of *information systems* and *computer-based information systems*.

What is the future for information systems? Nobody can predict it with any degree of certainty. As is well known, hindsight is always 20/20. Once someone solves a problem, the solution often looks obvious, and simply predicting the problem and its solution (in other words, the future), ignores the ever-so-true adage, “the devil is in the details.” With that caveat, where do I see the future in information systems heading?

What is likely is that certain existing trends will continue to prevail. Some yet undetermined technology will resolve or even eliminate some problems and open some other opportunities. Recall that 15 years ago, the World Wide Web did not exist! The need for data integration and the unified

searching it permits will intensify. Such integration will include text, image, audio and video data, and potentially some additional telemetric data. Integrated search of documents other than those available in convenient electronic form, for example, scanned documents, will become more prevalent. Today, each component of such complex documents such as signatures, hand-written annotations, logos, watermarks, stamps and text are processed separately. Higher accuracy in the search of such complex documents requires the integrated processing of these components, capitalizing on the nuances and strengths of one type of component to assist in the processing of the others. Cell phones are no longer phones; they are

now hand-held, computerized, mobile devices. However, their interface, both keyboards and screens, is limiting. Proper interfaces to these devices must be developed. Some of these interfaces will rely on locality and summarization-based applications. Users of these devices are often more interested in quick answers or short summaries of items in their current location rather than long lists of potential links to relevant resources.

In short, obtaining data is currently not and will certainly be even less of an issue in the future. Converting data into knowledge is the concern. How successful we are at deriving new insights from the data will determine how advanced our future information systems will become. ■