Information Services & Use 30 (2010) 99–105
DOI 10.3233/ISU-2010-0616
IOS Press

Emerging technologies to speed information access

Del Satterthwaite *
Director of Sales Engineering, Perfect Search Corporation

Abstract. Indexing and search technology has remained essentially unchanged since the 1980s. The industry still relies on these old technologies to access information. New and emerging technology is available that dramatically increase the speed, precision, and scalability of the index and search solutions we have today. This new innovation provides a core set of technology that can be leveraged by organizations to convert information into power.

Keywords: Search, index and query, performance, precision, scalability

1. Need for emerging technologies in indexing and search

The Gartner Magic Quadrant for Information Access in 2009 made the observation that search engine technology, the most mature part of information access, has not changed since the mid-1990s [2]. Most search engines use variations of indexing and search solutions based on old methods which include b-trees, hashing algorithms, inverted indexes, or bit-map indexes to retrieve data quickly. The core of search engine technology, whether it's used on structured data in databases or on full text in web-pages and documents, is the indexing and pattern matching algorithms that create data structures for fast access to relevant data. Others believe that Gartner may have been too kind. Analysts at International Data Corporation have suggested these core algorithms have been around even longer than that. These observations call attention to the fact that indexing and search technology, and the data structures required for fast information retrieval, have been around for at least two to three decades.

In an industry where product life cycles range from one to two years, and processor performance doubles every 18 months, it seems that indexing technology has lagged significantly behind. The question then becomes "Is there a need for faster methods of accessing information?". In Gartner's Magic Quadrant, most search vendors listed are not focused on developing a more efficient way to access information but on packaging the information to be more useful and understandable. While these improvements are helpful, what happens when consumers try to apply these solutions to the large data sets they have today and larger data sets they will have over the next several years? In many cases, search solutions are prototyped and show they are capable of finding information in the test corpus, but when they move into the mainstream, the search solution becomes unwieldy, unresponsive and incapable of handling the full data load. It's no wonder that Jane McConnell, in the Enterprise Search Sourcebook states that 3 of 5 customers are "not really satisfied" or "not satisfied at all" with their enterprise search solution [5].

*Address for correspondence: Del Satterthwaite, 400 West Dynix Drive, Provo, UT 84604, USA. Tel.: +1 801 450 2744; E-mail: del@perfectsearchcorp.com.

Jean Bedford in the same book states that typical Fortune 500 companies use at least five different search vendors in their organizations [3].

In 1982, John Naisbitt published *Megatrends*, in which he stated:

> We are drowning in information, but starved for knowledge. This level of information is clearly impossible to be handled by present means. Uncontrolled and unorganized information is no longer a resource in an information society, instead it becomes the enemy [6].

The book was published in 1982, and many may feel that this reference is outdated and inconsequential today. While it may be, can we really expect the core technology of index and search that were derived during that same time period, to be applied successfully to the problems of 2010? Using the technology of the 1980s and 1990s, while accepted because there have been no alternatives, only leaves us one option to rapidly access the data we have: throw more hardware at the problem. However, throwing hardware at the problem is not a long-term, viable solution. With data growth at the level we have today, (doubling every 18 months according to Gartner) and our thirst to understand it increasing, it's obvious that changes are necessary.

So to the question, "is there really a need for faster, more precise access to information?". The answer is clearly, "yes!". We have been running for 20 to 30 years on the same technology to help us find answers. Faster processors have helped, but the need for reconstructing the indexing model is undeniable.

2. New indexing and pattern matching technology is defined

In 2005, two friends met near Salt Lake City to discuss the need for better indexing and search technology to keep pace with the performance gains witnessed in the microprocessor. The two were Ron Millett, the original designer of search for WordPerfect and Novell, and Dr. Dillon Inouye, a successful business entrepreneur with a passion for engineering and mathematics. They were acutely aware that any gains in the index and search performance were a result of faster hardware and the underlying software had seen no significant change. They both had first-hand knowledge of the challenges that were associated with finding relevant information. During their discussion, Dillon recalled a recent visit to his parents' farm where he had observed a grain combine at work. "We need to eliminate quickly" Dillon explained, "the straw and chaff in these huge amounts of data, leaving small molecules of relevant information". Dillon believed that the metaphor of the combine could be applied to the problem of searching massive data sets with similar results by using hierarchal layers and multiple dimensions of speedups.

Pondering Dillon's idea, Ron envisioned how an earlier solution he'd developed could be expanded to another level. Ron later stated, "At that moment, I understood that focusing on elision of the irrelevant hierarchies, combined with a mathematical and semantic method, would enable the processing of the growing deluge of information at a fraction of the traditional overhead. Dillon and I both then realized that the efficiency of this approach could be a quantum improvement in both speed of indexing and speed of retrieval in a search system, and give higher quality results, using far less computer resources".

Although Dillon passed away a few years later, the direction and technology were set for this innovative approach to search.

The main tenets of this new technology are:

1. Dramatically different indexing and pattern matching technology that provides a much shorter path to relevant data.

2. Use of automatic elision to quickly eliminate large sections of the index that needs no further search.
3. Indexes that can reside on disk and still perform better than traditional memory intensive indexing and search systems.
4. Normalization of data so structured and unstructured data can co-exist in a single index and be queried at the same time with speed and accuracy.
5. Incremental indexing for easy ingestion of new data that can be searched immediately without sacrificing precision or relevance.

Perfect Search Corporation was founded based on the tenets listed above. A new product was introduced that challenges the old methods of index and search. It provides speed that is an order of magnitude faster than current technologies and precision that works equally well for database query as well as for full text query. It also provides scalability by using disk based indexes that allow for markedly larger datasets to be searched than traditional methods and incremental indexing methodologies that allow for quick additions of data to the corpus to be queried.

3. Case studies of the technology

Any new technology must be cost effective and have practical application in order to be economically viable and acceptable to the market. A study by Mary Brenner of the Wharton school suggests that disruptive, emerging technologies in established companies are being thwarted by Wall Street as they analyze the company's directions and use of capital [4]. Companies that pursue innovative ideas are deemed to be risking stable earnings for ideas that eventually may be untenable and prove a loss of resource in time and capital. Extensions or enhancements of old technology are deemed less risky and, as such, get better reviews by the Wall Street analysts.

It is rare that radical game-changing technology is produced by one of these large established companies due the scrutiny of the financial markets. Startup companies are inherently more risk tolerant and are more likely to introduce revolutionary technology to the market. The challenge they face is to prove the efficacy and value of their technology in established markets without a recognizable name or the pull of large marketing departments. These challenges are intensified when the claims of the new technology depart so radically from what is accepted in the industry. Such is the case with Perfect Search. One method to show value is to compare the new approach with more established products through benchmarks. Another is through sales, although initially it will be to companies willing to take a risk.

The following are examples of what has been done with Perfect Search technology.

Example 1. In a study with two database companies and Perfect Search, querying for the same information, using structured (80 million Social Security Death Index records) and unstructured data (3.5 million Patent records), performance was measured in the number of queries per second each technology could support using the same hardware. Figure 1 contains the results of tests using search software from Perfect Search, Oracle and Microsoft. The hardware platform was Dell R510 servers running Windows 2008 Server, with 32 GB of memory and eight 1 TB SATA disks. The test design, frequency and type of query were all based on actual customer logs for a production system.

The results demonstrate a dramatic increase in the queries per second using Perfect Search's new indexing and search technology. The results show a minimum of 10 times greater performance than the

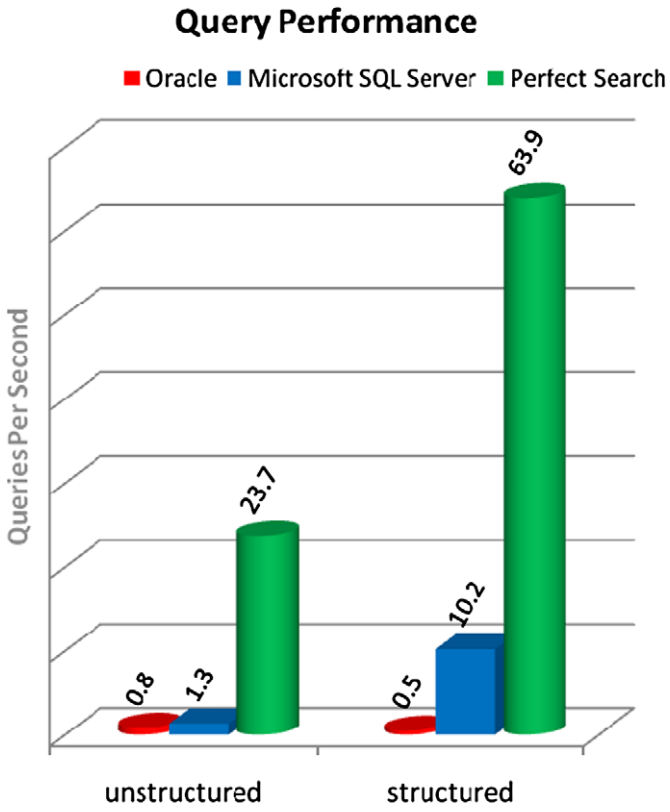

Fig. 1. Queries per second, Oracle, Microsoft and Perfect Search. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-2010-0616.)

traditional methods that are employed by Microsoft and Oracle. Although numbers were not captured during these tests, Perfect Search also used less memory, CPU and did fewer I/O read's in response to the queries. It should be noted that Microsoft and Oracle were not consulted to optimize the results in this test and could potentially provide better results given the opportunity. Perfect Search was also used "out of the box" and could also show improved results given time for testing and tuning.

Example 2. World Vital Records, Inc. was founded by Paul Allen, one of the original founders of one of the largest genealogy sites in the world, Ancestry.com [1]. As he was starting his new company he struggled to provide results to his customers in a timely fashion. World Vital Records was searching through 800 million records (60% structured data and 40% unstructured data) using the Open Source search tool, Lucene. Seven search servers were required to handle the load and were averaging 10 s per query response time. The response time ballooned to 40 s per query under heavy load. Using Perfect Search's emerging technology, the seven search servers were reduced to 1 with ample room for growth. Index time was reduced 100 fold, and the response time on all queries, no matter the load, is performed in less than one second. World Vital Records has grown since its initial implementation 2 years ago and currently has 1.6 billion documents being searched on one Windows 2003 server with 16 GB of memory and two 10 K-rpm disks. In Fig. 2, the results screen from a query, note the collection statistics for both names and databases searched.

Fig. 2. Collection statistics. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-2010-0616.)

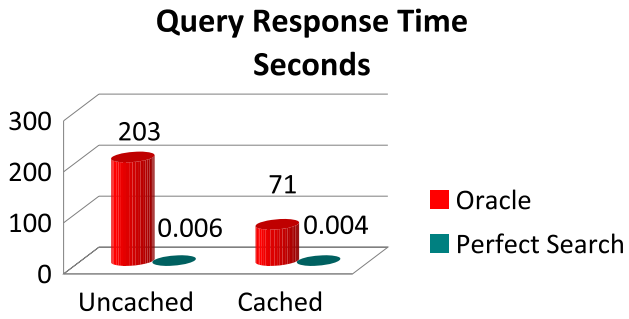

Fig. 3. Query response time, Oracle and Perfect Search. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-2010-0616.)

Example 3. A test corpus of 55 million records from a large library vendor was sent to Perfect Search for testing. The data came in Oracle export format and included scripts for indexing, and a set of test queries. Perfect Search loaded the data on a server, made a copy of the data for testing in Perfect Search and compared the results of Oracle and Perfect Search. In Fig. 3 we see the results of query response time on a date range query using Oracle Text style indexes. The results show the number of seconds it took to respond to a given date range query when done in a cached and un-cached environment.

With the wide discrepancy in the time to respond to the query, further investigation was done on the total number of I/O's required to provide the answer. In Fig. 4, the comparison of I/O reads from traditional indexing and search and Perfect Search's new method is shown very clearly. In this test, Oracle does over 1000× more I/O's to provide the same answer as Perfect Search. This indicates that Perfect Search is searching data in a revolutionary new way which dramatically reduces the hardware requirements for search systems.

Example (Breakthrough capabilities). NLP International uses patented MedLEE™ Natural Language Processing technology to for comparative analytics, predictive modeling, clinical research, automated

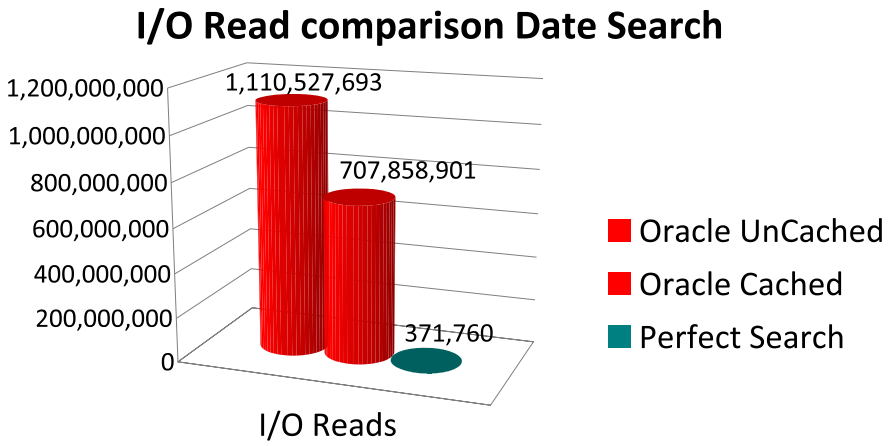

Fig. 4. I/O bytes read. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-2010-0616.)

adverse event identification, automated core measure identification, medication reconciliation and Semantic Search for healthcare and life sciences. MedLEE was developed over the past 20 years at Columbia University. MedLEE analyzes clinical free text sources, understand its meaning, encodes the data into standard ontologies and extract the information to drive decision support, research, coding and EMR population. Kyle Silvestro, VP of Corporate Strategy and Business Development at NLP International explained "MedLEE understand the clinical free text and provides a method to understand what is actually meant when a doctor enters 'no aspirin' or 'acute cirrhosis' in a patient record. The challenge was to create a solution that could leverage MedLEE's granular output, complex biomedical ontologies and enable clinicians to find complicated data very fast with NO IT Support. We had evaluated all other search technology and due to the nature of clinical data and the complexity of the MedLEE output we never found a correct fit, until they were introduced to Perfect Search Corporation". It quickly became apparent to Kyle Silvestro who has been following the market closely for the last 6 years; Perfect Search had a transformative technology and was a perfect fit for MedLEE. Together they have created MedLEE Search, the first Semantic Search and Retrieval Solution for Healthcare and Life Sciences. MedLEE Search has been called the Holy Grail for health care and will revolutionize the way we use clinical data.

4. Summary

Information is growing at a feverous pace. The sheer volume of that data can be overwhelming but the need to gain knowledge from that data is critical. The volume of the data we hold, and our desire for more exact queries in our search for what is relevant, have overwhelmed traditional indexing and search technology the industry uses today. Current indexing and pattern matching algorithms, the key to rapid index and search, have been unchanged since the 1980s or 1990s. We cannot continue to assume that we will be able to get the answers we need using old technology.

If we are to effectively process and query the volume of data currently being produced for the in-depth knowledge we require, we need new and better technology. Perfect Search Corporation provides an indexing and search solution that far surpasses the traditional technology dominating the market today.

Information is Power. Perfect Search's revolutionary technology provides unmatched performance, excellent precision and relevance, and unequaled scalability at lower cost. Its implementation promises to

more quickly and effectively uncover from data the relevant information businesses of all sizes demand, helping them harness the power that comes from knowledge.

References

[1] P. Allen, When free isn't free, available at: http://www.information-management.com/issues/20_4/-10018324-1.html.
[2] W. Andrews, Gartner Group, Magic quadrant for information access technology, RAS Core Research Note G00169927.
[3] J. Bedford, *Enterprise Sourcebook 2008*, p. 24, Information Today.
[4] J. Kirby, Wall Street is no friend to radical innovation, available at: http://hbr.org/2010/07/wall-street-is-no-friend-to-radical-innovation/ar/1.
[5] J. McConnell, *Enterprise Sourcebook 2008*, p. 28, Information Today.
[6] J. Naisbitt, *Megatrends. Ten New Directions Transforming Our Lives*, Warner Books, 1982.