A Beyond Search Report

Solving Search Problems with Perfect Search

A special report prepared by the Beyond Search team at Arnold Information Technology.

October 11, 2009, Version 1.6

Solving Search Problems with Perfect Search

Digital information poses more significant business challenges than information on paper. With Gutenberg's invention, a person could flip through books, file folders, and banker boxes of documents. With digital files, finding and discovering can be almost impossible.

> In the year 2002, print, film, magnetic, and optical storage media produced about five exabytes [about 1,000 petabytes] of new information to add to all of the existing information in the world. By the year 2010, we are on target to reach 988 exabytes. -"How Much Information?", University of California-Berkeley, 2003

The prediction from the experts at Berkeley was surprising to many in 2003. Today, an external one terabyte drive for a student's laptop costs less than $140. The price will fall below $100 within a matter of months. Organizations are awash in data. Some on servers. Some on laptops and netbooks. Other data on proprietary mainframe systems. There are audio files, video files, data in sprawling database tables. Then there is email. A large organization quickly generates as many as a billion (a one followed by nine zeros) in a single calendar year.

The problem is not new. And with the exception of a small number of information retrieval innovators, the problem has not been solved. A company in Orem, Utah, however, has taken a giant step forward. To understand the breakthrough, we need to take a quick journey back in time.

The Trigger

Two friends met over lunch about 45 minutes from the commercial center of Salt Lake City. Both men had participated in the early development of software search and often talked about search's inability to keep pace with the doubling of computing power every two years as commonly expressed by Moore's law. The two were Ron Millett, a 20-year veteran in search software, and Dr. Dillon Inouye (Ph.D. Psychology, Stanford), a successful business entrepreneur and professor with a passion for engineering and mathematics.

In early 2006, Dillon had often visited his parents' farm in a valley of the Wasatch Mountains because of their poor health. While on the farm, he had an epiphany as he watched a combine harvesting wheat. The combine was slicing and dicing away the straw and chaff, leaving just the wheat.

That day's discussion was about the problems associated with the growing mass of electronic information in organizations. Ron (now the chief scientist for Perfect Search in Orem, Utah) recalled how in that conversation, Dillon turned to him and said, “We need to eliminate quickly the straw and chaff in these huge amounts of data, leaving small molecules of relevant information.”

An Information Explosion

281 exabytes of information created in 2007. That’s 45 gigabytes (45,000 megabytes) per person worldwide

Rich media such as motion pictures, images, email, and RFID and other real time sources of information will boost the amount of information by a factor of six by 2011

More than half of the data will “not have a permanent home”; that is, **data will be “out there” in the cloud.**

Source: “The Diverse and Exploding Digital Universe,” IDC, 2008

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.

Now, pondering Dillon's idea, Ron saw how an earlier solution called hyperspace 1 could be enhanced to another level. Ron Millett said, "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 with a fraction of the traditional overhead. Dillon and I both then realized that the efficiency of this approach could be quantum improvement in both speed of indexing and also speed of retrieval in a search system giving a higher quality results an using far less computer resources."

Ron, a scientist who walks the narrow line between vision and blindness due to a medical condition, continued: "About two years later, Dillon died, and I continued the work, as he would have wished. I remember thinking during that original conversation: 'If we follow this approach to its logical conclusion, we can have a better and better search, a perfect search.' That's where the name of the company originated, and I believe that's what we have developed from his original insight."

Before diving into Perfect Search's technology, let's review the reality knowledge workers face.

Big Data, Bad Decisions, Uncontrollable Costs

Users today want a robust, high-speed, and flexible information access system. The range of user content and information demands are increasing. "Good enough" or adequate search is not enough. In addition, older architectures impose punishing hardware and infrastructure costs upon organizations. In today's business climate, economical operation goes hand in hand with:

- A need to locate the best documents or data to satisfy the user's information need
- A way to answer a complex question across structured and unstructured data
- A demand for search that goes beyond results lists with documents of less-than-close relevance to the user's query.

There are some significant limitations to traditional information access. Keyword search was once the high-water mark in search and retrieval. Language is ambiguous, and research indicates that users of keyword systems find it time-consuming to craft a query that "unlocks" the information treasure chest.

Pain for the Knowledge Centric Organization

One of the truisms of programmers is that bits are bits. Bits, as any business professional knows, may look alike, but they are very different and have distinct properties and value.

ITEM: Information in a traditional database provided by IBM, Microsoft, or Oracle, among others, requires specialists to unlock it. A simple question like "What uptick in sales took place with that discount we offered over the last five years for Mother's Day?" can take days or weeks before a manager has the information in a usable form. The person with the question must work with business analysts and other specialists. Even though the information is "in" the system, the manager cannot access it quickly. No one can accurately gauge the cost of such time delays.

ITEM: Unstructured information in the form of electronic mail and text messages clogs the in boxes of today's knowledge workers. In addition to the inefficiency and reliability problems of e-mail, the legal risk of these communications is skyrocketing. Consider a well-known vendor of software designed to perform eDiscovery for law-related business. The software developer used its own eDiscovery product in a legal matter, and the company was fined because its own product could not find certain e-mails related to the legal issue.

ITEM: Web information finds its way into organizations and into an organization's computers each day. Some of the information is of great value because it provides facts about the world outside of the organization. But when someone goes looking for such “harvested” web information, he or she may find it impossible to find what is needed and thus be forced to once again go out onto the web to perform research. Repeating what has already been done is very expensive. When you begin to consider cost of work hours spent repetitively on the same action, dollars can add up exponentially with each occurrence as work efficiency is compromised.

Key Words, Dead End for Busy Professionals

There are some significant limitations with traditional information access. Keyword search was once the standard in search and retrieval. But language is ambiguous, and research indicates that users of keyword systems find it time consuming to craft a query that zeroes in on relevant information. Consider this example of a newspaper sports story.

The story does not mention the word “baseball,” instead describing the game in terms of the actions of individual players. It is perfectly clear to the reader that the topic is baseball. But a search on the keyword “baseball” would miss the story, because the word “baseball” is nowhere in it.

Google is a keyword system that uses “voting”-user click behavior-to determine which document is most likely to be relevant to a person looking for information related to a given word or set of words.

Voting search methods don't work for organizations, because most documents are accessed infrequently or not at all. Yet when the need arises, the document must be located. Users of keyword systems find themselves struggling to come up with the “correct” keywords to find the “Smith invoice” or the “engineering proposal from last month”.

Some vendors have developed systems that assign documents to broad categories. The approach mirrors library classification: books on Shakespeare's plays are in the section labeled “PR.” The use of categories is handy for collections of like items such as books, but is less helpful for collections of unlike items such as invoices, customer e-mails, sales literature, and so on.

A small number of companies have developed systems that allow the user to formulate a query that can look in several databases at one time. The term used for such systems is federated search. The idea is that a user's query can be sent to three places: an index of e-mail, an index of proposals, and an index of purchase orders. The results may be presented in separate clusters (e-mails here, proposals here) or be merged into one results list. Regardless, the challenge in an organization is to keep the indexes current.

Variations on federation include producing a report for a user in which the results list is converted into “answers.” For certain types of information, turning results into a report works well. The challenge is to convert both structured (database) information and unstructured (for example, e-mail, PowerPoint, and Word documents) into a form that can be analyzed by the report generation system.

Studies by the Association of Image and Information Management, among others, point out that regardless of system, user dissatisfaction with search is in the 55 to 65 percent range. In short, regardless of approach, search systems are not meeting the needs of their users.

Federation: One System, All Necessary Information

Some vendors have described the federated search function as a portal delivering a roll-up of “federated content.” The typical federating method is to take a query and transform it into the syntax of other retrieval systems. Once the query has been translated, a federating system sends it to two or more other search systems. The results come back, are processed to remove duplicate items, and are sent to the user's browser for review.

Most federating systems do not build a local index of the processed content, nor do the systems support content repositories. Some systems can index certain content and then perform federation across the local index and other remote indexes accessible to the system.

The key point is that federating systems send queries to other systems. The approach saves the user the time and bother of sending a query to multiple systems one at a time. However, each system responds differently, and traditional federating systems may exhibit some sluggish response times.

This may be addressed by adding hardware and implementing more aggressive result caching. Vendors of traditional federated systems address delays using a variety of methods, including the use of parallel systems

The idea of searching many different sources, databases, and collections appears to be a good one on the surface, but search time delays are a problem. The problem is exacerbated by disappearing documents, sometimes called "broken connectors." An index from a remote system may point to a document that has been deleted or is no longer available, and users do not understand a link without the source content.

The principal hurdle for traditional federated search systems is cost. Using multiple servers and then scaling them when performance degrades is too expensive, too complicated, and too unpredictable for most organizations. As a result, search becomes a compromise.

Perfect Search and Hyper-Federation

Hyper-federation deals with the problem of information retrieval in a fundamentally new way. The idea behind hyper-federation is to tackle the problems of the amount of information, the different types of data, and the cost of infrastructure head on.

Hyper-federation processes content from multiple sources and creates a single, optimized index. Source data can be retained in a repository or in the search system itself. The low cost of storage and advanced content processing eliminates the need for time-consuming round trips between the user and the source.

Vortex: The Perfect Search Method

Perfect Search has developed a method that permits indexing multiple, sometimes geographically separate databases with a single search engine. The innovations at Perfect Search permit high-speed retrieval, drastically reduced hardware footprints, and extensive customization.

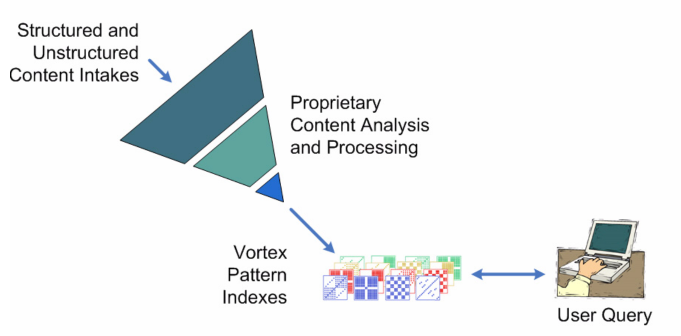
The table on the following page provides an overview of some of the differences between looking at information from different sources and Perfect's Search hyper-federation technique.

Perfect Search is one of the few organizations able to handle billions of documents and other instances of digital information using two or three servers instead of racks of cutting-edge devices.

Table 1: Traditional Federation and Hyper-federation

Attribute	Federated Search	Hyper-federation
Query Method	Query transformation and dependency on other systems' indices.	Query passed to a single hyper-index of structured, unstructured, and real time content.
Relevance	Approximate due to latency in remote system responses	Statistical and semantic methods with "just in time" refinement
Syntax	Free text, some Boolean operations	Fuzzy, Boolean, and free text queries supported
Response time	Dependent on remote system response and resources available to the federating system	Sub-second response time on low end servers
Index refresh	Dependent on remote systems	Indexes refresh on a licensee-stipulated schedule; for example, every 15 minutes, hourly, etc.
Content throughput	Dependent on remote systems	Gigabytes of text per hour on a basic server; more with additional server resources
Interface	Results list. Some systems include clustering and facets.	Customizable to licensee requirements

The diagram to the right provides a vastly simplified schematic of the Perfect Search technology. Several patent applications disclose the systems and methods invented by Dillon Inouye and Ron Millett, but the basic idea is hierarchical molecular representation instead of atomic indexing.



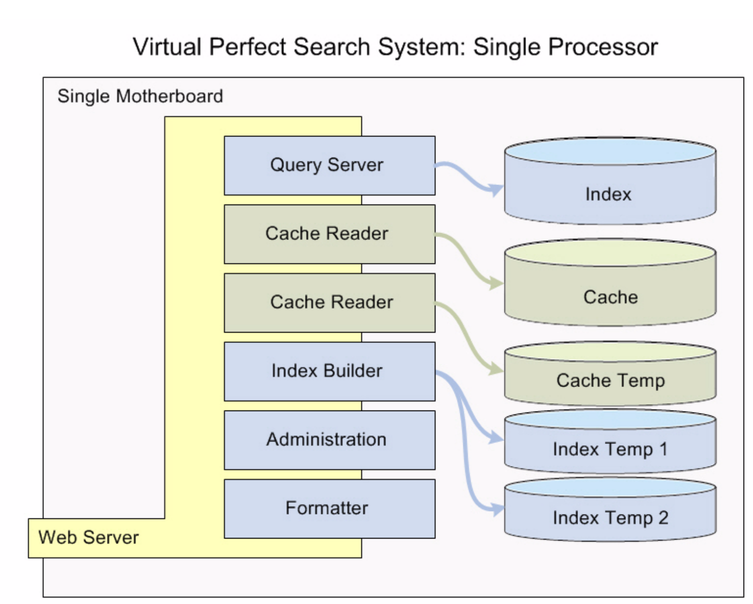
The triangular component represents the indexing process that converts information into molecules. These molecules are stored in pattern indexes. The user's query passes against these efficient indexes to return relevant, actionable results in milliseconds on commodity hardware. For more information, see "Pattern Index", US2007/016481, filed January 2006.

The system processes huge volumes of source information and reduces it to a hierarchy of indexes and accelerators without losing the sources' meaning. The system processes information into a representation of the content, ideas, and meaning of the source material in a series of small bundles of molecules arranged in "vortexes"-that means that each of these representations can "spin" a query to the precise information matching the user's information need. The system then "unwraps" the bundles of information molecules and delivers a relevant, on-target information answer. The efficiency allows the system to maintain performance in a disk-based implementation. As a result, one or two servers can do the work of six or eight servers required by other vendors' approaches.

The diagram below provides a vastly simplified schematic of the overall Perfect Search technology. The Perfect Search system ingests content in a wide range of file types and formats. The indexation process takes the "chunks" of content in the form of documents, emails, structured data sets, and Adobe Portable Document Format files, among others, and "distills" them. Key features of each content object are extracted, tagged, and written to the Perfect Search patented "vortex pattern indexes". These indexes are optimized and exposed in a way that makes low-latency retrieval possible. The indexes have two other important features as well.

First, if specialized data mined molecules for searching are used, the size of the Perfect Search index is much smaller than traditional search and retrieval systems. Larger indexes are also often used to fine tune exceptional performance on retrieval or to improve relevance rankings. However, even these larger indexes are created many times faster and also are accessed for retrievals many times faster than competing systems. Second, the structure of the index makes it possible to "slice and dice" information across the content types processed by the index. Cost and performance benefits are two benefits of the Perfect Search approach. Perhaps more important is the fact that the system makes it possible to have a single view of the content processed into vortex pattern indexes.

A quick look at the engineering used within the overall Perfect Search system provides a useful way to get a feeling for the computational efficiency of the Perfect Search system. The engineering developed by Perfect Search makes it possible to reduce the number of servers and storage devices typically required by traditional search, retrieval, and content discovery systems.



In the Perfect Search system, each server becomes a complete search system. The system obtains significant engineering, cost, and performance benefits because scaling requires "snapping in" another motherboard or server. The "administrator" is smart software that maximizes performance across the "virtual data center" the system implements.

The diagram to the left shows how Perfect Search creates complete "virtual" content processing systems within a single processor. What this means is that on motherboard with a single 64-bit CPU, the Perfect Search system implements an complete content processing systems. When two CPUs are utilized, the Perfect Search approach is to implement virtual, parllelized systems. These CPU implementations interconnect into an intelligent search grid. The breakthrough eliminates costs and, possibly more important in today's terascale information environments, latency in indexation, search, and retrieval.

The engineering approach makes scaling almost as easy as plugging in another server with multicore processors. The system automatically makes use of these resources. The pay off is that scaling is unburdened by the traditional configuration, tuning, and testing steps required with traditional search and retrieval systems. When coupled with the efficiency of the engineering of the vortex index, subsystems for the system, and the "virtual" innovation.

Several differences in the Perfect Search approach set it apart from the technology employed by other companies active in the enterprise search industry.

First, the efficiency of the methods makes it possible to generate indexes and keep the indexes current in near-real time. The throughput of a Perfect Search content processing system is measured in gigabytes of content per hour on a single commodity server.

Second, the vortex method eliminates the bottlenecks associated with passing a query against indexes in memory by minimizing the need for caching indexes in RAM (random access memory), thus reducing the cost of each server and delivering millisecond response time using commodity hard drives like those in a standard computer available at the local computer store.

Third, the flow of content through the indexing subsystem permits near instantaneous updates of the vortex indices. As a result, licensees of Perfect Search can specify the freshness of the index. E-mail, text messages, and news feeds become available with near-zero delay.

Perfect Search has filed for several patents on its approach. Other patent applications are in process. The numerical recipes, systems, and methods of the Perfect Search system boil down to three breakthroughs

First, Perfect Search processes content. The system can handle structured or unstructured content. The system analyzes the source information and generates a series of small-pattern molecular representations. These representations are hierarchical and dimensional. Their small size and proprietary structure make it possible to reduce the time required to process, update, and make use of information in the vortex pattern indexes.

Second, queries are passed against an abstraction of these pattern indexes. The metaphor of the vortex is apt. The system hooks into a concept at a high level and in milliseconds pulls matching and related content from the pattern hierarchies.

Third, the user can fluidly and intuitively interact with the system with near-zero delay.

Now let us look at a few of the innovations in Perfect Search:

Infrastructure: Perfect Search Does More with Fewer Servers

In order to take a query and pass it against multiple indexes or repositories, vendors have to have adequate servers, random access memory, bandwidth, and storage. Known issues with traditional approaches to information retrieval are boiled down to performance. As the volume of information to be processed goes up, so does the need for increased and faster infrastructure. The cost of infrastructure in today's financial climate can be difficult to justify. A number of companies have developed sophisticated methods to handle terabyte information flows, but these solutions are out of reach of most companies.

Power Consumption: Costs of Power and Cooling Sharply Reduced

Even a stable data center environment with reasonably up-to-date hardware faces the laws of physics. The technical and financial challenges of dealing with cooling and the power demands of servers, disc drives, and other devices are almost insurmountable in many organizations. One large corporation's systems manager revealed, “We can't get any more power in this location. It's a dead end.” The notion of running a “green” computing environment is a growing trend, and the constraints of heat and power--regardless of price--make it difficult, if not impossible, to keep pace with the volume of information in an organization. Power consumption for data centers per rack was one KW in 2000; it is now 10 KW per rack with plans to move to 20 KW; a major data center can require as much power as an aluminum smelting plant.

Latency: Sub Second Response for Source Data Display

Users expect documents to appear in a search systems output as soon as they are entered into the system. But when a document prepared in the morning cannot be located in the afternoon, employees are forced to find a workaround. These manual retrieval tasks slow down the work flow and add friction to the decision-making process. In a small organization, the delay and cost may be immaterial. But when an organization imposes manual retrieval upon tens or hundreds of employees, the costs of inefficiency soars. With the growing importance of fast access to e-mail and other key documents, a sluggish system may undermine or compromise a decision. Stale information is as risk laden as incorrect information.

Performance: Sub-second Query Processing under Heavy Query Loads

A user abandons a search process when it takes more than a few seconds to complete. In an organization, a sluggish system may be avoided. In one extreme case, a unionized environment permitted workers to run a query and then perform other tasks while the system processed the query for minutes on end. Not surprisingly, the investment in search did not increase user satisfaction or accelerate certain work processes. A fast system is one that encourages a user to interact in a fluid way with the information available. Many search systems lack the infrastructure to deliver sub-second response time. Complex de-duplication or clustering processes add to the sluggishness of a system.

Accurate, Flexible Retrieval: Taxonomy and Other Metatagging Supported

Traditional search and federating systems deliver acceptable precision and recall. Most systems today can display matching results. The issue becomes more complicated when a user cannot interact with a system to get a result quickly and when the system performs “background” operations that may inappropriately delete documents in an attempt to de-duplicate large result sets. The solution in wide use is to implement automated or semiautomatic methods of generating additional index entries for a document. Most users are not aware of advanced system functions, and the investment in semantic or other computationally burdensome indexing add-ons is wasted when they go unused or are under utilized.

Relevance: Multiple Methods

As Perfect Search molecules are processed, adding additional search terms often actually speeds up the search. This ability enables personalized and context sensitive “invisible” search terms to be included in the search. Term proximity, metadata fields, and key document paragraphs can all receive specialized relevance weighting. Word form and database field semantic analysis adds further relevance boosts. The rapid indexing operations enable data mining preprocessing. The system also makes use of document length and date (freshness). Users may enter one or more search terms or enter longer phrases or segments of text. Fuzzied queries for structured data are also supported.

Hyper-Federation in Action

Here's how Perfect Search uses hyper-federation:

1. Light-weight Hardware. Perfect Search uses mid-range servers such as standard two core eight processor systems with eight to 32 gigabytes of random access memory and standard desktop drives.

2. Perfect Search’s new content processing method. The breakthrough makes it unnecessary to rely on expensive caching, exotic high performance infrastructure, and fiber storage devices. Perfect Search's patented method makes use of advanced mathematical techniques to deliver sub-second content acquisition, index updating,

and query processing.

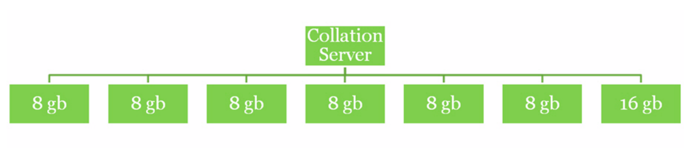
3. User interactivity becomes fluid because a query returns relevant results in less than one second when querying structured, unstructured, and real -time data such as e-mail or SMS text. Users spend less time getting information and more time applying it to their work.

Example 1: World Vital Records

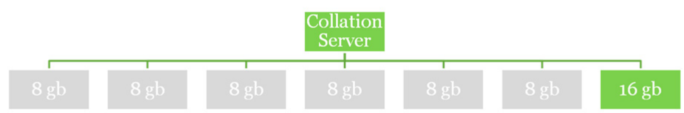
A database publisher found itself in a digital vice. In order to meet the online demands of its customers, more servers were needed. Adding more servers in a traditional, hosted environment boosted the cost of operations to an unsustainable level.

The company had to find a way to scale without the costs of additional hardware and power. The figure below shows the original configuration to handle 800,000,000 records with a capacity of seven queries per second. The hardware was 64 bit Windows servers with dual core CPUs.

Before The original configuration used eight servers and supported only seven queries per second. Index updating was measured in days even though staff added and modified records continuously.



After The Perfect Search solution used existing hardware and delivered 40 queries per second for 1.2 billion records across 12,000 separate databases with two servers. The other servers are available for expansion.



The cost savings have been dramatic, running at about 20 percent of the prior search system's. In addition to reducing query servers from seven to one, the company can grow without the burden of server capital costs and attendant maintenance fees.

The table below provides a snapshot of the going forward costs between the prior solution and the Perfect Search system. In this projection, the difference in the amount of hardware (20 servers vs. 1,160) is a huge savings in itself; that savings is increased as the slim down of architecture is passed on to the other costs inherent in operating the system such as utilities and maintenance. The largest savings would easily be found in reduction of work hours-much fewer are needed to operate the Perfect Search solution.

Table 2: Before Perfect Search, After Perfect Search

Cost Category	Perfect Search	Prior Search Solution
Servers	20 servers	1,160 servers
Server CAPEX	$60,000	$3,480,000
Recurring Costs (Power and Maintenance)	$40,000	$2,320,000

Example 2: Patent Documents

One of the most unwieldy types of content is patent applications. Most patent search systems require the user to enter search terms against specific fields. The Perfect Search implementation shown below in "Google Patent Search Results" service developed for ArnoldIT.com on a subset of the Google patent documents makes it easy for any professional to perform the search. The laborious entry of classifications and specialized tags has been replaced with a search box that supports free text, natural language, and Boolean statements. A single click displays the source document, in this implementation a full-text Portable Document Format file. Other display options are supported. Results may be customized or enhanced with taxonomy categories, if desired.

The Perfect Search patent demonstration is available at http://arnoldit.perfectsearchcorp.com/

A corpus or content domain can be enhanced with Perfect Search's taxonomy and tagging solution. In addition, Perfect Search offers a connector to integrate the Google Search Appliance using its OneBox API with the Perfect Search hyper-federating system.

Looking Forward

Perfect Search's technology represents a breakthrough in search and content processing for organizations. The company's approach can reduce the cost of deploying, operating, and maintaining a search system. The company can also enjoy more indefinable cost savings in man hours by the increase in efficiency developed with the Perfect Search solution.

The system can be deployed on premises or in a third-party data center. Content from traditional databases or unstructured sources can be indexed and made available with near-zero delay, not hours or days or weeks after the information has been placed into the system. The system allows a user to interact in near-real time, in a fluid, natural way to obtain answers to questions and iterate searches as a result of answers found. Third-party tools such as taxonomies, entity extraction, visualization, and text analytics can be integrated into the system.

Perfect Search works as a “snap in” to the Google Search Appliance and other enterprise software. Perfect Search's vortex method delivers speed, relevance, and cost effective scaling.

More information about Perfect Search may be obtained from the company's Web site at http://www.perfectsearchcorp.com.

Stephen E. Arnold, Beyond Search
October 11, 2009

Beyond Search Research is the research and marketing arm of Arnold Information Technology, a firm specializing in strategic advisory services. BSR provides technology and business analysis services to clients worldwide. The founder-Stephen E. Arnold-is the author of the first three editions of the Enterprise Report, three monographs about Google’s technology, and Beyond Search, a monograph about next generation search technologies, published by the Gilbane Group in April 2008. He with co-author Martin White wrote Success Enterprise Search Management for Galatea Press in the United Kingdom in January 2009. This special report was commissioned by Perfect Search in August 2009 and is the work of Jessica Bratcher, Donald Anderson, Anthony Safina, Stuart Schram IV, and Stephen E. Arnold. Mr. Arnold’s team publishes a Web log at http://www.arnoldit.com/wordpress which contains news and information about search and content processing. Information about ArnoldIT.com is available at http://www.arnoldit.com.