Search Engines: Evolution and Diffusion

Harry Collier, Managing Director, Infonortics, Ltd.
Stephen E. Arnold, Arnold Information Technology

Version 1.2, January 31, 2003

Preface

Harry Collier and Stephen E. Arnold have collaborated on projects for more than 20 years. This white paper is a result of a series of discussions held over the last year. Messrs. Collier and Arnold have watched the traditional search-and-retrieval services move from the center of the online world to its margins. At the same time, new search giants have emerged and redefined the meaning of the term *search* in significant ways.

Mr. Collier has a long and distinguished career in publishing, journalism, and online information. He is the founder of Infonortics, Ltd. (www.infonortics.com and located at Tetbury, Gloucestershire in England) a specialized information company that hosts an annual conference on search engines. Now in its eighth year, the conference is the preeminent event for the developers of advanced search engine technology and the meeting place for the leading providers of search technology and search services. Mr. Collier's influential analysis of the information industry and his broad industry influence give him a unique perspective from which to view a subject that has become one of the core functions of networks. Presentations from most of the Search Engine Meetings to-date can be viewed via a click at www.infonortics.com/searchengines.

Mr. Arnold has worked in many facets of the online information industry. He was one of the individuals closely associated with ABI/INFORM, Business Dateline, the General Business File, and dozens of other online products. He has worked on a range of search-related projects. These include assisting the U.S. government with the indexing of Federal government content and consulting with a number of "next-generation search" developers. His trilogy of monographs—*Internet 2000* (1994), *Publishing on the Internet* (1996), and the *New Trajectory of the Internet* (2000)—have tracked the business and technical impact of search-and-retrieval technology.

This white paper is designed to bring together for managers, systems professionals, and developers baseline information about search engines. It is offered without charge because the authors believe that the amount of misinformation and the growing confusion over search can be addressed by a clear, factual discussion of this important and often misunderstood and misrepresented topic.

The authors welcome reader comments and suggestions. Send them to editor@arnoldit.com.

Harry Collier, Tetbury, Gloucestershire
Stephen E. Arnold, Harrod's Creek, Kentucky
January 31, 2003

Search Engines: Evolution and Diffusion

1. Executive Summary

Table 1: The Main Points of This White Paper

1	Search is complex and often not precisely defined prior to undertaking discussions about technology or making major licensing decisions.
2	Search is a moving target. Innovations will come, but the main drivers in 2003 will be revenue.
3	The market for "pure search" is often seen as limitless. The reality is that search-centric firms are not likely to generate revenues over $200 million unless they have a "secret sauce" like Overture's pay-for-placement model.
4	Advanced search technology such as natural language processing are not yet ready for broad deployment when large domains of content must be processed in near real time.
5	True "innovation" in search is often incremental or a variation of "old wine in new bottles." Search systems must be selected on the basis of documented, well-defined customer needs and specific engineering requirements.

2. Introduction: The Challenge of Search

Most users know the frustrations of searching the web: No hits, or half a million. People want answers, not lists of where an answer might be.

Content is expanding more rapidly than indexers – humans, robots or hybrid systems – can index. The content inside most organizations is doubling every six to eight months, roughly the same pace at which the Web is expanding. Furthermore, when someone changes a document, the indexing for that document has to be updated. Change in a document set typically affects 20 to 30 percent of the documents regardless of when an index was last refreshed. Many documents become frozen while others are in a state of flux. Finally, the number of people looking for information continues to rise. As organizations expand their digital collections, more employees are looking for digital needles in 'digital haystacks' (a phrase use by Dr. Matthew Koll at the 2001 Search Engine Meeting in Boston, Mass. and available in his presentation at www.infonortics.com/searchengines/sh01/slides-01/sh01pro.html).

Second, making software do what humans do when they understand language is 'hard'. A series of rules that rapidly locate documents, analyze their contents, index the documents, place them in a useful category or conceptual pigeonhole, have yet to produce a economical, reliable, practical solution. Advanced technologies show considerable promise on small collections of documents; for example, research reports about cancer. However, when confronted with the informality and compression of electronic mail, a mix of popular magazine articles and newspaper articles and a collection of PowerPoint files with charts, graphs and executive jingoism, they fall over. Add a digitized radio program or a video and search engines are hamstrung. Humans are good at sizing up a cartoon, understanding a table of values and interpreting what a skilled writer means when he or she crafts a striking metaphor. So far, software is just not this intelligent.

Third, the cost of indexing is skyrocketing. Unless a skilled indexer works without pay, or for the hourly rate of a Kentucky Fried Chicken cook, a single index entry for an article, document or Web page can cost from $3 to $50 or more. The reason for the variance is the complexity of the content. Indexing technical materials can be time-consuming. Figuring out how to index a magazine article that talks about six or seven companies and their stand on a social issue consumes indexers' time. Proper nouns must be identified and normalized (there is a big difference to a computer between 'IBM' and 'I.B.M'). The person

looking for information is rarely sensitive to such nuances.

To build a commercially-viable intelligent search and retrieval system, costs are high. Infrastructure—the servers, telecommunications, and support systems—is expensive. In addition to machines, bandwidth and personnel, the time required to solve the problems associated with intelligent, automated indexing is a black hole. The burn rate for advanced search and retrieval systems is significant. The demise of such "successful search systems" as Excite, Northern Light, and Inktomi are reminders of the risks associated with search engines. The investments for a system such as Google's or FAST Search & Retrieval must measured in the tens of millions of dollars. Despite the costs, the complexity and the risks, entrepreneurs, scientists and financiers keep trying to break the back of the intelligent indexing problem.

3. The Hottest Search Trends

The bulk of this white paper talks about what might be called "classic search." For the degreed information professional, *classic* means Boolean queries passed against collections of content unified by an editorial policy. For the professional schooled in engineering, political science, or literature, *classic* means using online resources accessible via a Web browser. At some happy time well in the future, the two classics will come together. Now the views are polarized. The same terms are used by both groups but with subtle and sometimes not so subtle differences. For that reason, it is useful to define what one means by *search* before launching into a costly search-and-retrieval project. Search is a term bandied about too frequently and easily. The assumption that each person participating in a discussion of search has the same understanding of the concept is a very naive one.

Every two or three years, search sheds its cocoon and like a butterfly emerges in bright finery with a different look and feel. Much of what is "hot" or trendy in search is only slightly new. A bit of poking under the marketing promises, one finds string matching, thesauri, and statistical relevance ranking. That is not to say there is nothing new in search. There are some surprising developments that warrant careful consideration. The trends that seem to have momentum include:

A. Metasearch

A metasearch is a software system that takes the user's query and sends it to multiple collections of content or to different search engines. The query is passed against each of the "sources" and the results are returned to the user. The newest metasearch engines are remarkable beasts.

There are two broad types of metasearch systems. The first sends the query to Internet content, concatenates or integrates the "hits", and displays them to the user. The metasearch system performs this function from a standard Web pages. Examples of this type of metasearch may be found at Ixquick, Pandia, Ez2www, and Killer Info, among others.

The second uses a software client residing on the user's machine or on a server connected to the user's network. The client approach allows different functions to be executed on the user's machine, eliminating the delays that would be encountered if these functions were run remotely via a dial up modem. Examples of metasearch running on a client machine include Copernic, which boasts more than 500,000 users worldwide, and IntelliSeek Bulls Eye.

Metasearch is important for three reasons:

1. Many individuals looking for information know that no one source or Web index is likely to contain most of the information that is likely to shed light on the user's question. As more people become cautious with "hits" from search services that sell a

high ranking in a list of results, the need to look at multiple sources increases. Pay-for-placement is creating among some online searchers an increasing awareness that "hits" in a results list may be biased, skewed, or downright wrong.

2. With the uncertainty about copyright, a metasearch engine can incorporate for-fee sources and display them in a list of hits. If the user wants to view the full-text article, a fee can be charged for that document. Online users do not like the taxi-meter pricing model, but it does provide a measure of protection in the event of a copyright question.

3. The possibilities for creating false metadata are increasing, not decreasing, as Extensible Markup Language and display mechanisms based on XML become more widely used. A precise query can on some search systems return "hits" that are not directly related to the user's query. The Copernic metasearch tool provides some built-in features that provide some protection for this type of "hit" distortion.

4. Intranet searches—that is, searches run against content created by employees and stored on servers behind an organization's firewall—require metasearch technology. In most organizations, it is not possible to create one respository of content with one master index. There are for security and other reasons multiple repositories, multiple indexes, and often a large number of distributed systems. Although still in its infancy, metasearch technology is likely to play a more important role in Intranet search.

B. Pay for Placement

Pay for placement means that a Web site or content owner pays a service to deliver traffic to a specific Web site. The company credited with pioneering in this search segment is GoTo.com, now Overture Services Inc. It is by hundreds of millions of dollars and millions of searchers one of the most successful search companies in recent history. The company generates nearly $600 million per year in revenues and returns profits in the tens of millions of dollars in a down economy.

Google, the *doyenne* of search, has embraced pay for placement in two ways. First, Google sells in-line advertisements. When a user searches for *white papers*, the Google results displays advertisements with links to companies producing white papers. The second way is a near-clone of the Overture approach. The BBC site features the Google search engine. A search for *travel* returns BBC partner sites before other sites. The functionality to seed results with "hits" to paying customers or preferential partners resides within the Google architecture.

Litigation is pending between Overture and FindWhat.com. Overture has also filed suit against Google. In both cases, Overture alleges that its patent on pay for placement has been infringed. European pay for placement providers such as eSpotting have been emulating the Overture model with some success.

Because pay for placement delivers traffic, the era of biased results sets has officially begun. It is difficult to hide $600 million in revenue. However, most search companies overlook Overture's technology, its search and retrieval architecture, and its huge customer base at their peril. Where most advanced search companies starve for lack of customers, Overture's business has been robust.

C. Portals and Portlet Mania

Every organization wants a portal. More accurately, every organization wants a Web page to make data and applications available in one place so the time lost hunting for data and learning applications can be reduced. Search is a major component of portals. (A *portal* is a Web page doorway to different content and functions available via an organization's Intranet.) Search is implemented as a *portlet* function; that is, a separate application that appears on a Web page. When a person uses the "search" function on Yahoo!, for example, the search is a portlet or a mini-application that anyone can use without training.

Portals, like knowledge management and content management, can mean different things. The search function, however, is provided by the portal toolkit integrator. Commercial portals are often built on BEA Systems WebLogic, IBM WebSphere, Sun Microsystems Net ONE, or Microsoft Dot Net frameworks. Search software from Verity, Autonomy, PC Docs (Fulcrum) can be integrated into a portal. It is important to keep in mind that each of these portal companies offers search as a built-in function. These search services are designed to provide fairly simple functions, so upgrading the search is a common practice. The companies that have done the best job of licensing their technology for portals are Verity and Autonomy, which accounts for the two firms' combined market share of about 60 percent of the enterprise market. Google offers its "Google in a Box" product for portals. However, due to limitations in Google's index architecture and the difficulty of making enterprise sales, the Google in a Box product has not had a major impact on portal search at this time.

D. Peer-to-Peer Search

Peer-to-peer search or *P2P* as it is known in the popular press is an important innovation in search. P2P search freeware has contributed to the double digit decline in audio CD sales in the last two years. However, P2P search is extending its reach to other types of content, including video and text.

The idea behind P2P search is that software running on each user's machine can be used to prepare a listing of available content. The software from KaZaA, BearShare, and LimeWire—three of the popular P2P search tools—uses the Internet as a transport mechanism. Software looks for machines that have exposed content and matches a user's query to the content on these different, distributed machines. There is no single index such as one would find in the PC Docs Fulcrum product or in the typical indexes constructed by Inktomi, for example. (It is important to note that Verity and Autonomy, as well as other commercial search software support versions of P2P tailored for commercial Intranet use.)

The importance of P2P will grow, particularly with the advent of automatic aggregation software. Moreover, NewsNow, and hundreds of similar sites are examples of a combination of automatic aggregation, clustering, automatic indexing, and basic search functionality. A good indication of what's coming in peer-to-peer search can be seen on the Google news page (http://news.google.com) where real-time aggregation is combined with Google search.

E. Database Queries

Search is usually thought of as looking for articles and documents. In organizations, often the most useful information is held in structured databases. Database software from Oracle, Microsoft, and market-leader IBM come with search tools. However, the queries must be crafted using Structured Query Language. Consequently, average users cannot search the complete contents of a legacy accounting system for past purchase orders from a Web browser.

Progress is being made, and the database companies themselves are among the leaders. Microsoft is acknowledging that SQL (Structured Query Language) is difficult for many computer users. The company

is now shipping a more user-friendly query and report tool from Crystal Reports with its database products. Search companies such as Easy Ask offer products that can make the contents of databases available from a Web page. Verity's K2 product provides similar functionality.

Significant progress is being made in making database content more easily searchable. In 2003, enhanced functionality and more seamless integration of free-form document queries with the data retrieved from databases will find its way to market. However, for most organizations and most online searchers, database content will require a separate search. Results of the text query and the database query will have to be combined by the user. The first company to "solve" the problem of merging these two types of data will enjoy considerable commercial interest.

F. Other Trends

There are other trends that many will identify as more important than the four we have highlighted. Metadata is often identified by the consultants who follow search as important. We believe that any information about the information in a document should be usable by the search software. However, as noted, the advent of pay for placement means that search engines can intervene and ignore metadata or use it, depending on who writes the checks. Metadata is important, just not more important than the other trends we have identified. Some believe that video indexing and online text searching for specific scenes is a fast-growth area. We agree because non-text represents the largest percentage of digitized content available on networks. Audio and video data, for example, remains almost impossible to search without specialized tools like LimeWire or huge investments in systems to make sense of 30 frame-per-second video. Others include support for true cross-language searching (see, for example, Gregory T Grefenstette's examination of multilingual retrieval at www.infonortics.com/searchengines/sh01/slides-01/sh01pro.html). A query in English returns relevant results from sources in multiple languages with translation functions a keystroke away. The key point to bear in mind is that the integration of search results from audio, video, textual, and hybrid (text plus Excel spreadsheet data and a dynamic SQL database) across sources in different languages is the Holy Grail of search. For the foreseeable future, search usually means text. Search companies have a long way to go in that relatively well-defined domain. Digitized video is a challenge for the future.

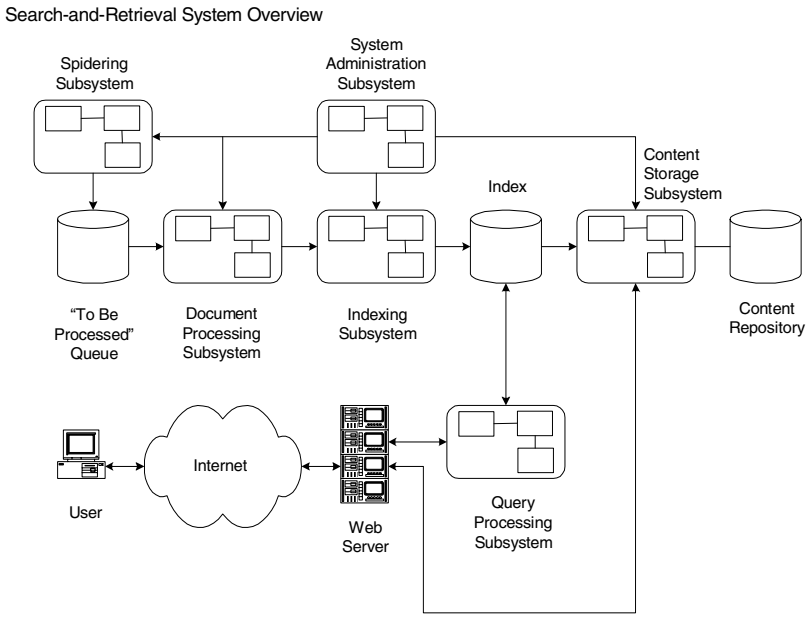
4. How Search Engines Work

The schematic below shows a simplified system architecture for a spider-based search and retrieval system dealing with a body of content (documents). The key elements which may be implemented in a different ways by different developers are:

- the search query itself. Other systems display results in graphical form. These indexes must be refreshed. Some search systems cannot refresh their index. The content must be respidered and reindexed. Other systems can process new data and incrementally update the index. Indexing is the key component of any search engine.

- The content storage subsystem is a feature of some newer search systems. A good example is Google's "cache" function. The document and often an HTML representation of a PowerPoint or PDF file are kept in the storage subsystem. The content respository is the area where the spidered source and its versions are kept. The content storage subsystem performs administrative and queuing functions when a document is requested.

Figure 1: Overview of a "Typical" Search-and-Retrieval System



- The spidering system. This is the series of scripts that visit a Web site, copy the content to a storage area where it is held until it is indexed, and perform a range of functions such as calculate a value so that on subsequent visits the spider can determine if new content has been added to the site, captures content to a specific depth on the site, and so on.

- The document processing system converts spidered content to a format the indexing system can process. For Intranet systems, the document processing system uses file conversion routines that make standard desktop software and legacy system files understandable to the indexing system. For Internet system, Hypertext Markup Language and Extensible Markup Language files are among the most commonly indexed document types. Google and FAST Search & Retrieval support Adobe Portable Document Format and common desktop application file types such as Microsoft Word and sometimes PowerPoint slide decks. Databases are, as a rule, not easily accommodated in text-centric systems. Spidering systems have constraints imposed by available bandwidth or network latency. The "to be processed queue" refers to storage devices that act as a storage area for data spidered but not yet processed or indexed.

- The indexing system reads the individual files and performs such functions as assigning or extracting words and phrases from the file, mapping metadata such as who authored the document and the date the file was saved to a storage device, and the file type of the original document and any other data stored and tagged as metadata.

- The index is "where the rubber meets the road." The query is passed against the search system's index. The index contains information about the documents matching the query. Indexing systems include a broad range of subsystems that can perform such tasks as calculating a relevancy score for each retrieved document, displaying a summary of each document, and other services. Advanced systems take a query in one language and return matches in the index from other languages. Other systems analyze a query and return documents that match the query but do not have any of the terms in

The query processing system performs a function that mirrors the indexing functions. The query is converted to statements that can be matched or passed to the index. Natural language systems take the user's query and "index" it. Boolean statements are converted to logical statements that are used to locate matching records based on entries in the index. Other systems match the words in a query to the words in an index.

- The Web server is a catch-all for the devices that display Web pages and allow user's to retrieve information from the index.

Some explanatory comments are warranted:

1. A 'natural language' service must be implemented in processes that require additional system resources: [a] parsing users' queries and [b] indexing the content so that linguistic, semantic, syntactic, inferential clues, clusters and categories (sometimes data to facilitate visual representations of data), proper nouns, bound phrases ("White House" "airline terminal") or complex nominals and other higher order 'meanings' can be extracted automatically from documents. Thus, 'subsystem' must be understood to be a robust computer and software infrastructure.

2. The content flow is constant and the indexing for certain documents must occur in 'near real time'. Delays of even a few minutes for certain types of information are not acceptable. Examples of content demanding no-delay indexing are intelligence reports in a crisis, and financial information for a person at a trading desk. The subsystems to handle the content functions are necessarily robust. The schedule for spidering, respidering for changes and spidering for new sources or Web sites, is continuous. Thus, the luxury of shutting down and processing content is not available. Few systems in the market today can do near real time indexing of content that requires iterative analysis. Near real time systems are associated with key word identification or symbol extraction associated with fast-breaking news stories or financial data. The costs associated with bandwidth, spidering subsystems and scripting for integrated and synchronized document processing and indexing and index updates are substantial.

3. The administrative services required to manage a near real time service that supports users with content delivery are difficult to implement quickly. Considerable fine tuning of the complex subsystems is required. Consequently costs associated with the management of the search and retrieval system are usually underestimated. These costs can be equivalent to the cost of building the document processing and indexing subsystems. Expenditures between $2-10 million for software, infrastructure, and administrative systems are not unknown.

4. Maintenance, enhancements and stability of the entire system are major challenges. Because of the need to run parallel functions and perform operations such as updating the master index while a query processor is performing a search, system management is a daunting task. A configuration management program and rigorous software engineering are necessary to keep the system from crashing or what search experts call "falling over" when any change is made. Most search systems are not fault tolerant, and they do fall over or crash. Advanced systems, particularly in their initial year or two in the marketplace, are not ready for prime time. A commercial environment is different from a controlled research environment. Datops SA (Paris) and Inktomi's search units went out of business because the overall systems were not economical to operate in a financially-constrained environment.

There are hundreds, if not thousands of search software companies offering their products. Of these, Verity, Inc. (Mountain View, California), Autonomy Ltd. (Cambridge, England), and Open Text, Inc. (Waterloo, Ontario), and a handful of others are profitable businesses.

Against this background, it is easy to see a select few companies are emerging as serious Web indexing utilities and providers of Intranet search. Companies with less robust systems will pursue customers with limited content to index, or a specific problem to solve such as document or knowledge management in a single enterprise. The technology of companies such as iPhrase, Inc.(Cambridge, Massachusetts) and Easy Ask (Littleton, Massachusetts) is very solid and best deployed in controlled environments where the volume and types of content can be more precisely defined.

What this means is that search and retrieval challenges must be met with tools that are right for each job. There is no one right way to search. Matching the technology to the content, and then to users' needs and available resources, remains the right way to think about search and retrieval.

5. Sizing the Market

Whoever finds the solution to search challenges will hit the financial jackpot. The ecology of information guarantees continued innovation and exploration of the boundaries of search and retrieval.Yahoo! has discovered the value of providing a fast, easy way to locate information on the Internet. Yahoo! has limitations, but it has a large user base because of the system of showing a searcher choices and operating by pointing-and-clicking. (This section refers to the table appearing in Annex A: Market Size.)

Only a system that allows a user to interact effortless with a search engine to get useful result is likely to have a good chance of attracting Yahoo!-level traffic. For specific niches such as financial services firms, a better mousetrap–what search gurus call *precision* and *recall*–when it comes to locating information offers a solid payoff.

The financial facts of search are highlighted in the table below. The data used are for the last four quarters for the publicly-traded firms.

Table 2: Financial Results of Six Search-Centric Firms

Company	Revenues (000)	Total Net Income (000)	Comment
Autonomy	$51.3	$7.3	The No. 2 has revenues of less than 10% of Overture's revenues
Google	$300.0	$50.0	Privately-held. Authors' estimate
Humming-bird Ltd.	$372.1	$4.9	About 20% are search related revenues
Open Text	$158.7	$23.3	About 30% are search related revenues
Overture	$569.2	$84.5	36.8% of unadjusted total. More than 60 percent of adjusted total
Verity	$96.0	$9.6	The market leader is one-sixth the size of Overture in terms of revenue
Total	$1,547.30	$179.60	In Annex A, the balance of the search industry generates. To adjust for non-search, the reader may wish to decrease the revenues by 80% for Hummingbird and by 70% for Open Text. Both firms sell other enterprise software and services which comprise the bulk of these firms' revenues.

The table on pages 14-15, updated to December 2002, provides a more granular view of the search and retrieval market than most overviews provide. Several points warrant highlighting:

- In the United States, the largest markets for search software are the Fortune 500 companies, the largest professional and financial services firms, and the U.S. government. The number of buyers in any of these segments appears large. The reason the pool of customers seems large for systems that can cost well into seven figures is that the same customer buys several search solutions. Search companies themselves find that Verity, Autonomy, Open Text, and Microsoft search will be running in different departments on different content domains. A portal vendor may implement another search instance as will a document management and enterprise software solution provider. The figures in the table have been adjusted to reflect an overlap of as much as 60-80 percent in certain corporate and governmental customer segments. One search does not fit all.

- The license fee for search is often negotiable. The reason is that software is a small part of the total cost associated with search. The professional services associated with search are often included in market size calculations. We believe that it is more useful to look at the cost of the license and then negotiate with the vendor or other software integrators for the specific work needed to get the system up and running. Search companies are becoming services companies.

- A handful of companies dominate search. These are Verity, Autonomy, Open Text, and Microsoft. Microsoft is often ignored in search market estimates. However, the company includes basic search functionalities in some form in its Office and server products. For this reason, low-end providers of search find themselves spending considerable time and money making sales in small- and mid-sized markets. A built-in Microsoft search is "good enough" for many users and, over time, will place increasing pressure on search companies offering standalone search.

What observations do these financial data support. First, the total available revenue from licensing search software in 2002 is estimated to be less than $2 billion from all market segments. The authors have 70 percent confidence in this estimate, which gives the search market in the United States a value in 2002 of less than $2.0 billion and not more than $3.0 billion. The majority of the revenue is accounted for by Overture, Verity, Autonomy, Google, Open Text, and Microsoft. Most smaller search companies are not likely to survive unless they make a major market breakthrough as Google and Overture did in the last three years.

Second, advanced search and retrieval, usually lumped in a broad and not too useful category of 'natural language processing', is driven or has been driven by military and intelligence initiatives. Intelligence agencies are struggling with complex problems in information management in a command-and-control environment with high stakes. Funding over the last decade has been strong due to the robust U.S. economy and perceived security threats from extremist organizations and hostile governments. The technology that many start ups are describing as 'state of the art' or 'next generation' is anywhere from two to four years behind what the advanced laboratories are now working on. The pipeline for systems that have been around for 24 months or longer is still a long one. Research and development dollars enter the pipeline and commercial products come out the other end, usually a few years after the project is no longer classified. Stated simply, many 'new' products are not new. Start ups commercializing software from the military, intelligence and law enforcement sectors require time and money to create a commercially viable product. The commercial product must run in a price-sensitive, stable, reliable environment. Research projects and government initiatives operate under different rules.

Third, a few companies – Verity, Convera (formerly Excalibur Technologies) and Autonomy – appear as examples of successful companies in multiple markets. In fact, the list of commercially viable search and retrieval companies is a short one. Even former industry leaders such as Fulcrum (now part of the Hummingbird PC Docs entity), Open Text, and DT Search are struggling to find customers, revenue and a

sustainable competitive advantage.

Fourth, non-American companies do not figure prominently in the table. There are a number of active non-North American search and retrieval initiatives. Countries with a position include France (Pertimm, Kartoo), Japan (Fujitsu, Justsystems, NEC), Israel (Media Access Technologies, Korda Technology) and Russia (Yandex), among others. At this time there are more than 700 separate non-North American Web spiders and directories. For a list maintained by the authors, view the pages at www.arnoldit.com/sitemap.html.

6. Scrambling for Dollars

Not surprisingly, the search and retrieval market space is a hotly contested one. The niche has high visibility. Somewhere between 65 and 80 percent of polled Internet users say search is the chief use of the Web. Search is the second most used Internet service. Electronic mail is the most used service.

Investors get an adrenaline surge when someone spins a tale of frustrated users (about 90 percent of 180 million users wanting an 'advanced search engine'). But most users do not use advanced search features. When given the option of typing a sentence or a word or two, 90 percent of the users opt for the one- or maybe two-word query. Yahoo! and America Online are popular because they make finding information easy. Ask Jeeves had a reported 334 million unique users in February 2000, prior to the Internet downturn. This contrasts sharply with Yahoo!'s billion plus users. If natural language is a home run, Ask Jeeves is lucky to get a turn at bat.

However, the financial performance of some of the best known search and retrieval companies has been lackluster, turning in below average or poor financial results. Therefore, within the last 12 months, companies with sophisticated search and retrieval technology have repositioned themselves. Table 3: Search Engine Repositioning below provides a snapshot of a number of 'strategic shifts' in direction as these companies strive to generate sustainable revenue.The information in the table warrants four observations:

1. None of the companies has been able to build a sustainable business with basic search software licensing regardless of the presence of advanced technology. A secret sauce, such as for-fee services or commissions on content licences, is needed to make cash flow.

2. The companies have repositioned themselves, abandoning markets where sales were too costly or too small. When one company abandons a segment, others enter it.

3. The technologies in this table have been generally known and available for more than a number of years. One must not underestimate the challenges of marketing certain advanced search-and-retrieval technologies.

4. Search-and-retrieval is not a gentle, easy business either for the marketer or the buyer.

Table 3: Search Engine Repositioning

Search Company	Search 'Tech'.	Old Positioning	New Positioning	Business Model
Applied Linguistics (formerly Oingo)	Linguistic process when text loaded with NLP-'light' front end	NLP tools	Ontologies for Intra-net content collec-tions and services	License software and provide ontology consulting

Search Company	Search 'Tech'.	Old Positioning	New Positioning	Business Model
Ask Jeeves	Acquired Teoma and replaced Direct Hit technology	Easy query	Enterprise search and associative search results based on Teoma technology	License 'engine' for enterprise portal. Charge for professional services and support.
Autonomy	Spider for Intranet indexing	Knowledge management	'portal in a box'	License engine for Intranet portal indexing and wireless device search tool
Brightplanet	Index and search structured databases	Web indexing	Access to content in structured databases	License engine and sell services to build text mining systems
Convera (formerly Excalibur Technologies)	Text, image, and video search	ASP service plus site license	Enterprise search and text mining	Obtain new investors and return to site license business model
divine Interventures (formerly Retrieval Technologies and Northern Light)	Noun tagging, result clustering, and Web indexing	Indexed and filtered news feeds and Web indexing	None	No viable business model
HNC Software	NLP and noun extraction	Intelligence tool	Health care and enterprise intelligence	License software to organizations; fees for customization
iPhrase Inc.	NLP processing tuned for content domains	Web content indexing	Enterprise and Web content indexing	License software and provide professional services
Open Text	SGML DB	Web search and Intranet indexing	Enterprise applications, including knowledge management and collaboration	License tools to e-commerce sites wanting collaboration, database, services, and search in a one-stop shop
PLS / AOL	Web and Intranet indexing using probabilistic algorithms	"Find a needle in a haystack"	No cost Open Source software	None. Out of the search and retrieval business
Verity	Topics technology which puts content in categories	Enterprise search and OEM deals with other software products requiring search	Enterprise search, including access to structured databases	Text and SQL search licenses, OEM deals, plus professional services and maintenance
Yahoo!	Manual record creation with pass through to spidered index	Web directory of popular sites	Acquire Inktomi and shift to for-fee directory listings and Inktomi-generated Web search	Shift to for-fee advertising and subscription model

7. A Closer Look at the Types of Search and Retrieval Systems

The gulf between 'consumer NLP' in Ask Jeeves and cutting-edge technology from MIT's Advanced Computing Center, is wide. As a result, one cannot use an umbrella term such as 'natural language' and have it make any sense without providing a context for the term.

Table 4: Useful Search Concepts provides a snapshot of the principal commercialisation avenues search and retrieval companies are following at this time. Several observations may be useful as a prelude to scrutinizing the data in the table:

1. A search and retrieval engine can be repositioned by marketers with the addition of one or two adjectives. There is a short distance between search and the woolly 'knowledge management' market, for example.

2. The market is broad enough to support search and retrieval systems that are at opposite ends of the technology spectrum. Autonomy's statistical approach and the intelligent software in Vivisimo are fundamentally different. In search, whoever gets to a person with a need, and makes a case that a particular solution will 'work', is going to get the business. Technology is not always the deciding factor in search. There is a fair amount of technology sleight of hand in search.

3. A number of companies are following in the footsteps of Direct Hit (a popularity engine) and Google (a link analysis engine). Research by Andrew Tomkins at IBM Almaden Research Center in San Jose suggests that only about one-third of Internet sites are strongly connected. This means that a smaller index of popular sites and the sites that have the most links will perform reasonably well. Google has expanded its indexing and link analysis to cover several billion sites. However, the actual number of public Web sites is unknown. The top 100 Web sites, regardless of whose scorecard one uses, get the majority of the clicks. The difference in number of users between the top site and the 25th site is measured in millions of users. There may be billions Web pages, but 99 percent of them retrieve modest, if any, traffic. Popularity means that highly specialized sites may be difficult to get indexed and, therefore, find.

Table 4: Useful Search Concepts

Category	Comment	Example
Utility spiders	These services index the Web and Intranets for a fee. Inktomi charges $150,000 to set up a Web spider and then $0.02, or whatever can be negotiated, for each time an item is viewed.	www.inktomi.com drives www.alltheweb.com www.google.com
Rules based query	The user types a query as a sentence, question, or string of words. The parser matches the user's question with hand-built templates. Ask Jeeves has recently added popularity searching, a directory based on the human-built Open Directory and a product comparison service	www.easyask.com www.databeacon.com
Human based query	The user types a query as a sentence, question, or string of words, and a human expert answers the question.	
Clustering service	Recursive algorithms put documents that are similar together. Depending upon the algorithm, the clusters can be large or more finely divided. Computationally intensive process restricts clustering to subsets or collections of documents.	www.inxight.com www.stratify.com

Web spiders	Services that find a Web site, copy its pages, parse the pages, and make the index searchable to users. Excite has added link analysis to increase 'precision'.	www.altavista.com www.google.com www.alltheweb.com
Pay for placement	Users buy a listing on the first page of hits for a particular key work or phrase	www.overture.com www.espotting.com
Spiders with Boolean and free text query	The user can enter a word, a phrase, or a Boolean search statement.	www.lexis.com ww.altavista.com
Human-built directories	These directories use the input from Web users to build directories or points to Web sites.	www.opendirectory.org www.about.com
Manually constructed thesaurus	The search engine indexes only the terms in a user-edited or constructed thesaurus.	www.dialog.com www.verity.com
Metasearch	An engine submits a query to multiple other engines.	www.ixquick.com[1] www.mama.com www.metacrawler.com
Distributed shared architecture	Information is posted by the person who created it or who owns it. Napster-like technology allows other users to know information is available and to access it via a 'virtual content switchboard'.	www.limewire.com www.bearshare.com
Geospatial	The user points a wireless device in a direction. The system displays a menu of businesses in that direction	www.anarcti.ca www.kartoo.com
Mathematical approach	System makes no reference to 'meaning'. Indexing, search and retrieval are derived mathematically from a language system that cannot be understood by humans; for example, the speech of dolphins or the 'meaning' of one statement based on previous statements of the same class.	www.autonomy.com[2]
Natural language	The user can type a sentence, question, terms, or cut-and-paste a block of text. The engine parses the query and matches the query to the index.	www.iphrase.com
Spiders and human built directories and indexes	Queries may be selected from a taxonomy or a word, phrase, or string of terms can be entered. The 'hit' list points to sites selected by humans first and then displays sites indexed by a spider.	www.yahoo.com www.looksmart.com
Link analysis	Users may search entering words, phrases, or strings of terms. The results are returned based on either the number of links pointing to a site or on the number of clicks in a time interval a site garners.	www.ixquick.com www.google.com[3]
WAP centric	Spiders index specific sites, convert data to WAP format and create a searchable index accessible by pointing-and-clicking a mobile device's keypad.	www.pinpoint.com

1. *Ixquick uses algorithms derived from index fund evaluation. In addition, Ixquick incorporates both link analysis and click-rates to determine the relevance or importance of a hit in the metacrawler results.*
2. *While not a search engine as the concept is used in this document, this company is pioneering in the use of predictive, 'fuzzified' statistics. The initial application is in predicting the needs of Web users at a particular site at a particular point in time.*

3. *Google has begun to accept advertising. Since January 2000, the company has signed up more than 30 firms. The link analysis views a link from Page X to Page Y as 'one vote'. The importance of a page is determined by the number of votes a page receives. The page that casts the vote is analysed to see if it gets votes. Votes by pages receiving votes are more important than votes by pages receiving no votes. Google assigns a Page Rank score based on votes and importance of the voting pages.*

8. Natural Language Processing in 60 Seconds

University computing research laboratories are the petri dishes for NLP, ranging from well-known facilities such as Syracuse University and the University of California-Berkeley to some lesser-known facilities run by individuals who shun the glare of public relations and venture funding for advanced research. One excellent example is Dr. Edward Fox, head of Virginia Polytechnic Institute and State University's advanced text laboratory. Other innovators come from non-academic backgrounds. A good example is the former broker and index fund expert who developed the Ixquick metasearch engine working with some friends on Manhattan's lower east side.

Within the last ten years, much of the work has been driven by funding from the Defense Advanced Research Projects Agency and other government initiatives. The reason is anchored in the changing nature of security, defence and warfare. Digital warfare requires fast response. Sending paper memoranda has never worked. Today, the need to query computer systems as rapidly and fluidly as possible is as important as it ever has been.

The thrusts of the funding in the last five years have been reflected in the types of search and retrieval systems (of which NLP is a subset) that have attempted to become viable businesses. Specialists can rightly argue that there are many more 'categories', but for the purpose of this white paper, the categories in Table 5: Types of Search provide a useful way to get our hands on a slippery subject. Reasonably well-defined categories include:

Table 5: Types of Search

Retrieval Category	Definition	Examples
Statistical	These are called probabilistic systems. The indexing system keeps track of how many times a word like 'farm' is present in a collection of documents. The documents with more 'farm' words in them have something to do with farms, farming, farmers, etc. Word counts drive the indexing and relevance components of the system. There are various 'tricks' employed to widen or narrow what is retrieved, usually by thesauri or some type of look up table.	www.pls.com[1] www.autonomy.com
String matching "plus"	The system compares user input strings with strings in documents and uses proprietary techniques to provide what the user wants.	www.thunderstone.com[2] www.verity.com[3]
Linguistic processing	A series of modules examine a document, to locate paragraphs, sentences, and phrases. The system then uses look up tables, rules, and algorithms to determine the relation of the elements to the document as a whole. These systems are recursive and computationally intensive. (The number of calculations required to perform recursive algorithms is substantial.) Cost limits the application of linguistic routines to domains with a modest number of documents that do not require frequent updating.	www.iphrase.com

Boolean	Named after the mathematician George Boole, Boolean syntax specifies what must be present in a record or document for it to be retrieved. The problem is that the user has to know how to form a Boolean statement in words. This is a level of effort most users are not willing to take. Many NLP systems take a user's input and convert it to a Boolean statement. The 'NLP' part of the system is an interface issue.	www.dialog.com www.fulcrum.com www.opentext.com
Taxonomy systems	The user looks at a list of topics and clicks on the one that looks relevant. Each click displays either more choices or a list of 'hits'.	www.stratify.com www.clearforest.com[4]

1. *This system from Personal Library Software was one of the first commercially successful implementations of Cornell University's Dr. Salton's algorithms. America Online bought PLS to index chat. After the routines were integrated into the AOL environment, PLS was placed in the Open Source software collection. It is free and it works.*
2. *This is the search engine used by eBay. Thunderstone licenses stemming tools to America Online and other developers. EPI Thunderstone is based in Cleveland, Ohio and has an excellent product but a low profile in the text retrieval industry.*
3. *This is a client-side metacrawler. The user can type in words, phrases or sentences. The system looks for strings and then launches a search against about a dozen of the Web indexes; for example, www.all-theweb.com.*
4. *ClearForest builds a taxonomy by inspecting documents and creating metadata. The company uses routines that it describes as 'linguistic' and 'syntactic'.*

Marketing messages for search suggest that each company's system performs every search and retrieval function at the highest level of precision and recall. That cannot be true.

Equally misleading is 'real time' indexing. Operating at *near real time* means that systems must be updated continuously. All processes must be fully parallelized and synchronized, otherwise expensive subsystems sit idle. Just like road entrances at rush hour, all work tasks must be gated; otherwise bottlenecks are inevitable. The changes or 'deltas' must be identified and the old stories removed from the index and replaced in near real time. Doing two potentially conflicting processes in the same 'live', in-memory data structure is tricky. The larger the index, the more the task shifts from tricky to impossible. Index updating may require swapping complete index structures and updating one while the other is available online.

Wire feeds must be archived and de-duplicated, which is a difficult task with nearly identical stories with metadata attached. Unstructured stories pose a different set of computational problems which can be solved by iterative processes, but these take time to execute. There is then a difficult problem associated with providing an index to the current and stories with a newer version on the wire.

This list can be extended. Our point is that *time* is required to handle a number of mundane and difficult tasks. The *asymmetry problem* says that the time required for these tasks plus the natural language indexing routines is greater than the time available. There are, of course, solutions to this asymmetry problem:

1. Scale the search and retrieval system dynamically. It is not possible to 'throw' new systems at a search and retrieval problem. System-wide scaling is necessary.

2. Invent faster algorithms. While theoretically possible, the impracticality of what the authors cals 'the Bell Laboratories' approach is well known. It is difficult to mandate solutions or inventions. At this time we are saddled with known programming languages and their limitations.

3. Buy faster, cheaper computers. Hardware lovers endorse this option with enthusiasm. However, faster hardware can only do so much to solve the problem of asymmetric

systems. When the 'problem' is volume of information and iterative processes that are interdependent, hardware provides minimal relief from what is a fundamental architectural problem. LexiQuest and other NLP systems are centralized and subsystems are interdependent. (For LexiQuest see, for example, Bernard Normier at http://www.infonortics.com/searchengines/sh01/slides-01/sh01pro.html). Real time operations and interactions among complex subsystems is not a hardware problem; it is a *complexity* problem. Unexpected events occur as a consequence of system interactions. Not only does hardware not solve these problems, but hardware can increase problems in sometimes surprising ways.

4. Process less data. This is the solution that virtually all of the NLP systems (and commercial database publishers, we might add) embrace. By gating the volume of information indexed and available for searching, the system can be made to work. The performance may be sparkling within the limits set by the content. The problem is that searchers are not too keen on searching just "some of the data". What is not included may have relevance to the searcher's interest, so multiple searches must be run on non-NLP systems. Usage of NLP systems often decreases over time because NLP does not deliver solid benefits to the user; namely, better results from multiple databases in less time. NLP creates the need to run more queries to find content not in the NLP or advanced search system.

5. Use word lists (thesauri), skip changes, ignore duplicates and update only a handful of records. Most users are none the wiser. These shortcuts are widely practiced in all search systems, not just those calling themselves 'NLP engines'. The computer scientists and computational linguistics experts express dismay when shortcuts are the only way around the asymmetry problem. At the end of the day, the short cuts are put in place. Practicality usually takes precedence over a system that does not work or is so slow no one will sit through a sales presentation. (Selling a slow, unstable system is a bit of a challenge.)

The bottomline. Exercise caution when licensing NLP search or any 'advanced' search service for that matter.

9. Innovation: The One Constant in Text Retrieval

Most text retrieval researchers are unaware of innovations now moving to market from France, Russia, Japan, Australia and elsewhere. It is highly unlikely that NLP will provide a fast payoff unless its use is clearly focused. However, innovation in NLP is inevitable for the following reasons:

1. Cookbooks of algorithms are widely available worldwide.

2. Cheap processors are abundant.

3. Mathematicians recognize immediately that search and retrieval is a mathematical problem. Advanced search is a problem. Ergo: mathematicians will apply their expertise to the search 'problem'.

4. Readily available bandwidth and computer resources encourage experimentation in metasearch, visualization of result lists, value-added reports instead of lists of possibly relevant material, and smarter and more subtle pay-for-placement functions.

5. XML documents and database content encourage innovation in retrieving data from

SQL databases and text files, then presenting that data in a single display.

What is the single major innovation that must be pursued?

In our view, the P2P technology warrants additional attention. Close behind are searches across information in multiple languages. Visualization captures the attention of some researchers who blend the graphics of the video game with relevance and clustering. It is difficult, therefore, to narrow innovation to a single search 'problem'.

10. Some Search and Retrieval Realities: A Summary

Advanced technology – even the best technology – does not assure [a] positive cash flow, [b] good search results, and [c] market dominance. Advanced search and retrieval in general (and natural language processing or visualization in general) are easy to talk about. The mathematical concepts underlying even relatively trivial (by today's standards) engines such as Ask Jeeves are interesting and entertaining to discuss. But there are some hard facts in text retrieval. The major one drives the rapid repositioning under way at many search and retrieval companies.

Which company has the 'best' search? The reality is that no simple answer is possible. At this time, the Verity, Autonomy, Hummingbird, Open Text, and Convera search and retrieval systems are stable, scalable and effective. That does not mean search and retrieval in general and advanced techniques such as NLP in particular is at a dead end. The company with cash and a desire to push the envelope in search should invest in two, possibly three different 'advanced' approaches. This is certainly the logic behind Sequoia's investment in Google and IPhrase. Smart money tries to reduce risk while increasing the payoff on the upside.

Most companies will want to license software from a company with a track record. The reason is that the systems have to work and pay for themselves. There is one point for search entrepreneurs to weigh: companies and organizations who want a 'good enough' solution can use an off-the-shelf utility, or free search and retrieval software.

Who buys into the advanced search and retrieval technologies? The customers for the high-end systems have not changed in the last five years: the intelligence agencies of developed nations, financial services firms, blue-chip consultancies, pharmaceutical companies and a handful of information centric enterprises such as Microsoft, Intel, and Pharmacia.

The outlook for advanced search is bright. The need for voice input to systems is great, and pieces of advanced technology such as NLP technology will be important. Thus, for the foreseeable future, there will be a range of search and retrieval solutions, continued innovation in the technologies required to provide advanced search, and a handful of people who are in the right place at the right time when the 'next big thing' hits.

While the 'mobile explosion' helps drive voice recognition, the payoff for advanced search is not as clear. The reason is that the vast majority of online users are content to type one, two, or three words, hit the enter key, and see what they find. After all, the excitement of hyperlinking and finding the unexpected is one of the fundamental drivers of the Web experience. As Internet usage grows larger, change will necessarily come more slowly, if at all, in this aspect of user behavior.

There is a need for search engines must be able to handle non-text objects, including pictures, audio and digitized video. Basic image and song location are not the answer. Larger and larger volumes of non-text data are flowing through the Internet, most not indexed. In theory, the Extensible Mark-up Language provides a way for developers and content producers to add metadata to non-text objects. Widely used,

XML can help index certain content types. Multiple languages must be handled. Translation systems today are crude but can produce satisfactory results on tightly structured documents such as scientific and technical papers.

Today computer scientists and computational linguists work to create a system that can analyze documents and other objects, understand what they are about, and create an index and abstract of each object's content. This work will continue for many years.

Harry Collier, Infonortics, Ltd.
Stephen E. Arnold, Arnold Information Technology

Annex A: Market Size

Niche	U.S. Customers		License Fee Range per Year	Market Size (000)	Selected Companies in the Segment	Comments
	Overlap among sets is 60%	Lower Boundary	Higher Boundary			Potential customer pool is smaller due to overlap; that is, same customer buys multiple search solutions.
Audio search	20,000,000	0	$10,000	$0	KaZaA, LimeWire	Not yet monetized
Branded search for Internet portals	5,000	$10,000	$1,500,000	$50,000,000	Google, Yahoo!, FAST Search & Retrieval	Search utility business has consolidated
Document management	1,000	$100,000	$3,000,000	$100,000,000	Documentum, FileNet, Optika	These companies license search functionality. They don't create it.
Lotus Notes centric	20,000	$0	$25,000	$0	Grapevine, Mondosoft	A marginalized business
Lower-cost shinkwrapped solutions	1,000,000	$1,000	$1,000	$1,000,000,000	DT Search, Integrated Digital Systems	Segment difficult to reach due to cost of marketing
Military-law enforcement	1,000	$50,000	$3,000,000	$50,000,000	Stratify, Dienekis Information Systems	Target for advanced search engines
Natural language processing	1,000	$100,000	$3,000,000	$100,000,000	iPhrase Inc., Phrasys Ltd.	Computationally intensive and limited to high-end or special purpose solutions
OEM components	5,000	$10,000	$100,000	$50,000,000	Verity, Microsoft, Pertimm	Verity is the market leader. Microsoft will gain strength.
Ontology	10,000	$25,000	$1,000,000	$250,000,000	Applied Semantics, ClearForest	Smaller search engines are repositioning themselves as software to produce metatags
Open source	10,000	$0	$0	$0	grep, Lucene	Not yet monetized
Pay-for-traffic search	1,000,000	$50	$250,000	$50,000,000	Overture, FindWhat	The money making segment of search: match queries to advertisers' Web sites
ASP site indexing	100,000	$0	$24,000	$0	Blossom, FreeFind.com	Remote services indexes Web site and provides a link to the index for that site
Standalone search for portals or special purpose indexing	20,000	$5,000	$2,000,000	$100,000,000	Google, FAST Search & Retrieval	Google, Alta Vista, FAST Search & Retrieval
Still image retrieval	10,000	$10,000	$1,500,000	$100,000,000	Convera, Apple Computer	Convera, piXlogic
Toolkits for enterprise search	10,000	$200	$2,500,000	$2,000,000	Verity, Microsoft	
Visualization	2,000	$25,000	$3,000,000	$50,000,000	Plumb Design, Anarti.ca, Kartoo	
Video image retrieval	5,000	$200,000	$2,500,000	$1,000,000,000	Hitachi, NCR	
Totals	2,200,000			$2,902,000,000		The lower boundary was used for the market size estimate
Adjusted for overlap	1,320,000			$1,741,200,000		Totals adjusted by 60%