TABLE 1: SEARCH IN THREE CONTENT MARKET SEGMENTS

Selected Attributes	Internet	Intranet	Special Domains
Point-and-click installation	No tuning and coding required.	Lower-cost or ASP services have simpler installations. More robust tools require customized installation and setup.	Custom work required. Most vendors support file types covered in Stellent's "Outside In" tool via license or by custom code.
Automatic indexing	Standard feature. Usually statistical approach to eliminate recursive calculations for linguistic tools. Some engines support external knowledge bases.	Varies by vendor. Dedicated thesauri and classifications may be required.	Training or tuning required. Vendors may have to customize engine to handle certain content types.
Classification of content	A feature of certain "newer" engines, e.g., Vivisimo.	Varies by vendor. Third-party tools required depending upon customer requirements.	Custom work required.
Internet file types[1]	Support for HTML and XML "standard." FAST and Google support Word, PowerPoint, PDF, and a handful other file types.	HTML, XML, Microsoft Office file types, plus common legacy file types, e.g., Rich Text format, Word Perfect, etc.[2]	Special filters and customer work may be required
Index refresh mechanism	Controlled by scripts.	Lower-cost packages provide minimal controls. More robust packages provide extensive controls.	Lower-cost packages provide minimal controls. More robust packages provide extensive controls.
Content removal	Re-indexing may be required. Manual intervention required.	Tools vary by vendor. Most rely on manual intervention.	Customer development required if file type is not directly supported by search package.
Firewall functions	Commercial systems are tightly engineered to protect service. Spiders generally observe conventions of robot.txt file.	Varies by vendor. Some engines cannot index through firewalls and update indexes without customization.	Varies by vendor. Some engines cannot index through firewalls and update indexes without customization.
Spidering depth	Script controlled.	Varies by search engine vendor. "Depth" and "security flags" often require customization to ensure that sensitive content does not "leak" into the more generalized service.	Controlled by custom scripts.
Security controls	Robust protection of the core system.	Security assumed to be a function of the intranet system, not the search system.	Controlled by custom scripts
Cross-server indexing	Supported. Some servers can be polled more frequently to maintain appearance of index freshness.	Varies by vendor. Support may require customization of the search system.	Controlled by custom scripts.
Cross-domain indexing	Supported. May require separate indexes as is the case with Google.	Varies by vendor. The "behind the firewall" index is often separate from the indexes of third-party content and "outside the firewall" Web sites. Complex issue affecting perfor-mance, security, and index freshness.	Custom code required depending on the content, domain, and content locations.
Support for Lotus Notes and Microsoft Exchange[3]	Not generally provided.	Varies by vendor. May require a third-party prod-uct or custom code to provide a single search box to access the Notes and Exchange content.	Custom code required if data are to be accessible from a search box. Third-party products often used to handle special domains, e.g., Data Beacon for data mining queries.
Graphical interface	Not included. Interfaces coded by search vendor or customers	Templates provided. Customization required.	Customization or third party products required.
Selective Dissemination of Information functions	Varies by user's level of access to the search functions. Not generally offered for "free" searching except in the form of My Yahoo! or Google News which are variants of SDI technology.	Varies by vendor. If not included, a third-party tool such as BEA Systems WebLogic or WebSphere can be used to provide the function.	Customization required.
User customizable inter-face	Application Programming Interface (API) provided.	Templates can be changed by customer. API may support third-party services for visualization of results	Customization required.
Field search	Limited.	Varies. More robust packages support Boolean and field searching.	Customization required.
Support for SQL database content	Not generally available. Development underway at major vendors and services.	Varies. Vendors offering SQL support may require one search box for "text" and one for "data."	Certain content domains may require a separate log on authentication and search process. Customization required.
Multilingual support[4]	Most vendors provide support for multiple languages. More support coming but Arabic and Chinese continue to lag behind Roman-alphabet language support.	Low-cost tools may have no support for a language other than English. More robust tools provide greater support. API allows integration of third-party services.	Customization required. Varies by data source. A database may require only row and field names to be translated; data are numeric and can be used as is.
Network load controls	Sophisticated and controlled through custom scripts or Web forms permitting values to be entered to control spidering threads and other load-centric functions	Varies. More robust packages provide control via scripts or Web forms. Lower-cost or ASP services may offer no or limited controls.	Custom integration required.
Training	Varies. Focus is on selling managed services, not training clients	Varies. Even low-cost packages focus on up-selling service and support "bundles."	Depends on how the client approaches the problem of special domain content.
System administration	Extensive tools relying on scripts and Web forms for most frequently "tweaked" values.	Varies by vendor. More common is a two-tier approach. The client can handle basic functions like which folder to spider and when. More advanced services are part of the support "bundle."	Depends on how the client approaches the problem of special domain content.
Branded content	Biggest vendors can process branded content. Billing and rights management are issues.[5]		

Footnotes

1 Wireless content poses special challenges not generally addressed by the high-visibility vendors of search. Specialized vendors such as Pinpoint in Durham, North Carolina, do focus on this segment.

2 A list of the file types supported by the "Outside In" technology now owned by Stellent, Inc. is available at http://www.stellent.com. Some search engine vendors create their own filters in order to avoid paying license fees to a third party.

3 Both IBM and Microsoft offer search software to handle content in these proprietary software environ-ments. The next release of MS Office will, for example, perform more robust searches of attachments for elec-tronic mail.

4 At this time, FAST Search & Retrieval and Google do the best job of supporting non-English searching of the public Internet. Their intranet customers can use these services. Customization is typically required to meet the intranet customers' needs. Pertimm, a French search engine, is one of the few products that supports a query in English against multilingual content returning hits across the languages in the corpus.

5 Copernic, a Canadian vendor of intranet and personal search software, is working to sign up publishers and integrate the content into lists of Copernic "hits." This service will become available sometime in 2003. Branded content is not the challenge. The hurdles are keeping track of usage, billing, and reducing the risk of unauthorized reuse.

TABLE 2: SNAPSHOT OF KEY PLAYERS

Company	Snapshot	Secret Sauce
Applied Semantics Inc. http://www.applied semantics.com	Originally Oingo, this Los Angeles-based company offers automatic classification and human-edited "ontology" services.	Company has found a lucrative market applying its technology to suggesting new domain names. See Register.com for an example.
Autonomy Ltd http://www.autonomy.com	Originally based in Cambridge, England, Autonomy has become the poster child for the European software agency. With Verity, one of the dominant intranet indexing engines.	Made Bayesian algorithms the solution to Intranet search. New initiatives include search and retrieval of audio voice-mail messages.
ClearForest Corp. http://www.clear-forest.com	A sophisticated classification and indexing engine. The product is aimed at corporations and intelligence agency applications. Plan on a six-figure price tag.	Features automatic bound-phrase extraction.
Overture Services, Inc. http://overture.com	Formerly GoTo, this service incorporates string matching and a number of other sophisticated technologies. The company has transformed search by monetizing the hits displayed based on who buys a word.	The company generates revenues that are roughly six times the revenue of Verity. The financial winner in search. Overture will be increasingly challenged by Google's listing business.
Pertimm SA http://www.pertimm.com	A product of French scientists, the Pertimm engine provides a suite of technology that supports Web services and handles queries in one language across content in any of the dozen languages the engine supports.	Software returns hits based on automatic query expansion and point-and-click navigation of "glimpses" or relevant extracts from a corpus.
Stratify, Inc. http://www.stratify.com	Funded by the U.S. government, Stratify performs a range of classification and indexing functions.	Positioned as one of the first "content discovery" tools. Human-assisted indexing added when software-only solutions need "tweaking."
Verity http://www.ver-ity.com	Verity, with Autonomy, holds the lion's share of the U.S. search-and-retrieval business. Customers include Adobe and metasearch provider Bull's Eye.	Owns Inktomi's intranet customers and the Ultraseek text retrieval engine used to provide search and retrieval for Bitpipe.com
Yahoo! http://www.yahoo.com	Yahoo! has shifted from a directory service, although Web site owners are encouraged to pay for listings to a search model. Yahoo! left Inktomi for Google and in 2002 bought Inktomi.	Inktomi provides custom Web spidering that can be costly to scale and refresh.

TABLE 3: SEARCH AND RETRIEVAL'S CHALLENGES

Function	Comment
Harmonization	The search engine must recognize, convert, index, and provide pointers to multiple types of PDF files, file formats, database files, etc.
Auto indexing	Novices and experts can search the index using terms common to their training and experience and find comparable results.
Clustering	The system groups like objects that meet the needs of lay people and professionals alike.
Learning-centric	The system monitors users' actions and adapts to those patterns in order to return optimal results for each particular user.
Administrative	The system manager interacts with the search and retrieval subsystems via interfaces that make explicit the consequences of settings and minimizes or eliminates the need for programming.
Scaling	The system can handle the number and type of documents it is asked to index and make subsystems avail-able without reengineering, when thresholds are crossed.
Distributed	The search system is modular and can be distributed to make the best use of available resources and to minimize adverse effects on system response time from heavy indexing.
Explainable	When the search systems make a decision about placing an object in a cluster, an audit trail or some other type of concrete explanation for the action taken must be available.
Change-aware	The search system and its subsystems must be able to acquire new content, recognize changes to existing content, identify available but unchanged content, and index the objects accordingly.
Date-aware	The search system must be able to handle various types of date and time information and use each in the appropriate context for a particular query; specifically, date and time stamp assigned by the system, file creation date, file change date, and implicit dates extracted from the content cues in the object.
Adaptive	As new content objects are discovered, the search system is able to identify the object, intelligently process the object, or notify the system administrator of the new object and request guidance for handling it.
Language	The search system can recognize languages, index them, support a query in the user's language, return results from any object in the corpus in either the original language or in the user's language. The user makes a decision and the search system behaves the way a particular user requires.
Trainable	The search system supports input and guidance from humans via interfaces explicitly designed to accept existing terms, new concepts, ontologies, taxonomies, dictionaries, thesauri, or knowledge bases as required.
Application Programming Interfaces	The search system provides a documented set of APIs or hooks so that authorized users can make use of the search system or a particular subsystem from another program or in support of another process.
Security	The search system integrates seamlessly with existing security systems, so that results intended for a user with a particular level of access display only content matching that access level.
Usage tracking	The search system provides a native usage tracking subsystem giving detailed information on a cycle set by the system administration. No third-party tools are needed to determine usage patterns.
Multi-object	The search system can handle database, text, and proprietary file forms in a structured, flat, or compound form.
Query flexible	The user can query the system using a single term, bound phrase, or free-text entry. These functions may be exposed by the system administrator making use of portlets (tiny prewritten routines) that activate a particular search-and-retrieval function.
Point-and-click	The search system should generate a Yahoo!-style directory when the system administrator activates that function. An administrative interface (as noted in this table) allows modification or "training" of the search system to handle certain objects in a manner specified by the administrator.