Filters are program that allow transforming a document format into another. From instance, web browsers understand the HTML text format anf the GIF picture format. It is useful if the server knows how to transform any text processor format into HTML and any graphics format into GIF. Some servers allow you to add your own filters.

See also "client / server".

Z39.50

(WAIS, Fulcrum)

Z39.50 is a national standard defining a protocol for computer-to-computer information retrieval. Z39.50 makes it possible for a user in one system to search and retrieve information from other computer systems (that have also implemented Z39.50) without knowing the search syntax that is used by those other systems. Z39.50 is an American National Standard that was originally approved by the National Information Standards Organization (NISO) in 1988.

See http://lcweb.loc.gov/z3950/

Customizable Stopwords

(WAIS)

A stopword is a frequently used word that, when encountered in a user question, is ignored. For example, since the word "the" commonly appears throughout the English language, it does not help distinguish between documents. Thus it is typically regarded as a stopword. Indexers includes a list of approximately 300 standard stopwords. Some indexers allow the publishers to customize the stopword list.

Structured Query Language (SQL)

(Oracle, Fulcrum, Architext)

A language that enables you to create and operate on relational databases. SQL is an ANSI standard. Example: SELECT * FROM Customers WHERE rating = 100;

Relational database / Fielded search

(Oracle, WAIS, Fulcrum, Architext)

A set of related informations stored in tables. Relational databases are often used to store numerical and factual data. More recently, they have been used to organized text, such a email and notes (e.g. Lotus Notes). Documents that contain fields such as email (subject field, date field, from field) are amenable to be treated by/as a relational database. Indexes by fields can be built. Example of natural language field search query: all emails of isabelle from May 1995 until now

Query report

(WAIS, Fulcrum)

A report generate by the search engine for a given query which may include: * The database being questioned, * The original question, * The Boolean equivalent or SQL equivalent of the question. * The number of documents and the number of words in the database, * The number of unique words in the database, * The number of times each word in the query occurred in the database, * The expanded search words, * The number of documents found that satisfied the question, * The amount of elapsed time it took to perform the search.

Personal notes / Annotations

(none that I know of)

The capability of annotating documents the you access on-line and to have your notes transferred to the new versions of the document automatically.

Filters

(WAIS, Fulcrum, Verity, Architext)

Index Overhead

(need to be benchmarked)

The size of the database index relative to the size of the source text files. e.g. Indexes are less than 15% of the size of the documents being indexed for Architext, $25 - 35\%$ for Fulcrum and up to 100% for WAIS.

Index Speed

(need to be benchmarked)

The quantity of data being indexed per hour. e.g. The Architext search engine indexes up to 300MB/hour on an SGI Indigo.

Retrieval Speed

(need to be benchmarked)

Speed of document retrieval for a given query.

Client / server

(all)

The data are accessed by clients (e.g. netscape) from servers (e.g. a Web server or a WAIS server). This is a particularly safe way of accessing data because the server can restrict the access for: personal use only, a restricted set of users, corporate-wide use or public use.

See also "security" and "billing / activity log".

Inverted File

(all)

For each word listed in the dictionary, the inverted file lists all the documents containing that word. In same systems, the inverted file keeps track of the position of the word in the document.

Incremental Indexing

(WAIS, Architext)

Incremental indexing allows you to add, modify, or delete documents without reindexing the entire database and without suspending user service. Incremental indexing modifies the index to reflect data changes since the last time the data was indexed. This capability is especially important for network publishers whose data changes often, and whose database size is large.

Boolean Operators

(Oracle, PLS, WAIS, Verity, Fulcrum, Architext)

Traditional way to search with the AND, OR, and NOT connectors. Example: telephone NOT cordless AND AT&T

Proximity / Adjacency

(PLS, WAIS, Fulcrum)

Proximity (within N words) or adjacency of two keywords. Example: ADJ (Chernov, bounds) searches for keywords "Chernov" and "bounds" when they are adjacent. ADJ (Chernov, bounds, 1) searches for keywords "Chernov" and "bounds" when they are adjacent or separated by 1 word, including "Chernov like bounds".

Wildcard Search / Right or Left truncation

(PLS, Fulcrum)

Single and multiple wildcard characters allow searching for word variants. Wildcards can be used within or even at the beginning of a word. Example: data* maps to "data" and "database"

Hit Highlighting

(PLS, Fulcrum)

Search words in retrieved documents are highlighted for easy identification of relevant text.

Automatic Subject Grouping

(Architext)

Automatic Subject Grouping allows users to view the results of a query grouped by topic and subject; this gives users an alternate way of seeing their data that complements a standard relevance-ranked list.

Concurrent Update

(PLS)

Databases can be searched and updated simultaneously. Upon successful completion of indexing, the new data is automatically available to the user.

Two-Phase Commit and Transaction Control

(PLS)

Ensures that an abort during an index update will not affect database integrity.

Query-by-theme / Concept searching

(Oracle, PLS, WAIS, Architext)

"Query-by-theme" and "concept searching" identify the strongest overall themes in documents, using grammar, meaning (Thesaurus lookup) and/or pragmatic relationships between words. They discover words, phrase, or query that are conceptually related (eventually dynamically). With this feature, you can find documents that do not happen to contain any of the words you thought to search with, but are nonetheless relevant.

See also "query expansion" and "relevance ranking".

Topic agents

(Verity)

"Topic agents" are programs that filter on-line of off-line documents relevant to a particular topic. In that sense "Query-by-theme" and "concept searching" are particular cases of "topic agents". "Topic agents" may be used to gather off-line information of interest and warn a user of the existence of such information, for instance by sending him email.

See also "query expansion" and "relevance ranking".

Advisors

(PLS)

Similar to "query expansion" advisors propose to replace words by close relatives but do not send them directly to the search engine. The user is provided with a list of candidates he can choose from.

See also "query expansion" and "relevance ranking".

Query by Example / Relevance Feedback

(PLS, WAIS, Verity, Architext)

A "query by example" consists in naming relevant documents in a query form to search for other similar documents. The server analyzes the submitted documents and determines which of the words in the documents are useful discriminators. It then uses those words to find other documents which share them. One of the primary uses of query by example is to help users quickly focus their search without the need to learn complex query languages. "Relevance feedback" is the ability to select a document or a portion from the list of documents resulting from a first search to find a set of documents related to it.

Numeric Range Search

(Oracle, PLS)

You can perform range searches for numeric quantities like dates.

See also “distributed database model”, “remote index”.

Relevance Ranking

(PLS, WAIS, Fulcrum, Architext)

Documents are retrieved and displayed, with relevance scores, based on their likely relevance to your request. This is a proven, powerful technique that gives you leverage for dealing with large, complex information resources. All results from a distributed search are merged and displayed in a single relevance-ranked list. The relevance may be based on simple word density (how many times a word from your query is found in a document), but may include words weights in your query and/or in the text, proximity relationships between words and more complex relations).

See also “query expansion”.

Query Expansion

(WAIS, Oracle, PLS, Verity)

A set of methods for enlarging the search domain of a given query to go beyond simple word frequencies. Proximities are established between the words in the query and other words, based on spelling, grammar, meaning, context, etc.

See also “advisors”, “stemming”, “fuzzy search” “query-by-theme”, “concept searching” and “topic agents”.

Stemming

(WAIS, Verity)

“Stemming” is a technique used to automatically derive variations of a queried word. These variations are then used as part of the search. If stemming is used, then when a data set is indexed, word stems are indexed where possible. For example, ”dancing,” ”danced,” and ”dancer” would all be indexed as ”dance.” A question containing the word ”dancer”, would then turn up documents that may also include ”dance”, ”danced”, and ”dancing”. Two types of stemming are supported: Plural and Porter stemming. Plural stemming attempts to determine the singular form of a word. Porter stemming attempts to find the real base, or stem, of a word and derive any possible alternate variations. The stemming algorithm is selected by the database administrator prior to indexing the database.

See also “query expansion” and “relevance ranking”.

Fuzzy Search

(PLS)

“Fuzzy search” minimizes the consequences of bad spelling, typographical mistakes, and OCR errors in a database. Words with similar spelling are considered as well as the original word.

See also “query expansion” and “relevance ranking”.

Authentication features help support subscription-based online services. Some engines are also equipped with billing options that allow for customized prices to be set either per database, per document, or per number of characters downloaded. The activity log may keep track of:

- Total Number Of Connections
- Number of Different Machines Connecting
- Total Number of Searches
- Total Connect Time
- Total Search Time
- Searches Returning Zero Hits
- Total Number of Documents Retrieved
- Total Number of Databases Searched
- Number of Searches with no Database Name
- Number of Searches Requesting Help
- Average Number of Seed Words per Search
- Number of Searches using Relevance Feedback
- Number of Server Warnings
- Number of Server Errors
- Number of Search and Retrieval Requests for each database
- List of Client Machines
- Client Information

See also "client / server" and "security".

Natural Language Searching (NLS)

(Oracle, PLS, WAIS, Fulcrum, Architext)

You can enter a search using normal prosaic construction of words or sentences. Multiple languages may be supported.

Master index / Back-of-book index

(Oracle, PLS)

A "master index" system identifies important piece of information in a document. MasterIndex can be used to:

* Create a back-of-book style index for a document–listing concepts, definitions, and other information types that may suggest keywords to users

* Create hierarchical indexes for querying and browsing through groups of documents

* Populate databases with structured content information, including actions (i.e., who did what to whom)

* Create hypertext links for related information in a document and among multiple documents

* Enable intelligent information-extraction agents to track specific information and trends

Note: if the master index is written in a standard form, the authors can provide it along with their data, which makes the global index richer.

Hypertext Navigation Links

(Oracle, Architext)

Automatic creating of hypertext navigation links.

Information Agents

(Oracle, Verity)

Programs that track specific information and trends.

Distributed Database Model / Multinode Server Searching / Virtual Databases

(Oracle, TRW, PLS, Architext, Fulcrum)

A retrieval engine following the distributed database model allows users to transparently navigate through multiple databases regardless of their location. Both documents and indexes can be distributed. The distributed database model offers three major advantages: * Publisher control of 'look-and-feel', pricing, and editorial policies. * Seamless full text searching across local and remote servers. * Database groups that provide unique, personalized views of information. For example, a Market Research database group could consist of twenty separate databases stored on three different Web servers around the world. Some of the information is for internal eyes only, while other content is useful marketing information designed for customers. When users execute searches against the Market Research database group, they will be transparently searching across some or all of the twenty disparate databases as if they were one.

See also "master index".

Security: User Authentication / Encryption / Kerberos

(PLS, WAIS, Fulcrum)

Initial authentication is controlled by the server in the form of a user login ID and a password and further authentication at the level of the database group, restricting users to certain database groups, depending on authorization. Public key encryption can be used to encrypt documents and/or authentications. One of the popular public key encryption mechanism uses the fact that it is difficult to factorize large numbers into prime number factors. If A wants to receive information from B, he sends to B a public key which is the product of two prime numbers. B sends back a message encrypted with an algorithm that uses the key. Only A can decrypt the message because the decryption algorithm requires to know the prime factors of the key. Kerberos is an authentication system designed by MIT for use on the Internet which offers centralized key management for database holdings.

See also "client / server".

Billing / Activity Log

(PLS, WAIS)

SGML

Standard Generalized Markup Language. This is a standard for describing markup languages. HTML is an instance of SGML.

HTML

(supported by all)

HyperText Markup Language. A collection of tags used to mark blocks of text and assign them particular meanings. For instance, a word or phrase in a text can be associated with another document (hypertext link). HTML is used by the World Wide Web. A Web browser (e.g. Netscape) interprets HTML. A hypertext link may be revealed by underlining the word. A mouse click on the word will result in opening the associated document.

World Wide Web Robots, Wanderers, and Spiders

(none are spider)

World Wide Web Robots, Wanderers, and Spiders are all names for programs that traverse the Web automatically. They follow HTML hypertext links recursively to index large quantities of documents.

See also "remote documents".

Remote Documents / remote index

(PLS, Verity)

Some database administration tools can index documents that are on a remote server that is not running a copy of the main search engine. These documents remain at the remote site until they satisfy a query, at which point they appear in a hitlist with a URL link to the remote server; when the link is activated, the document is retrieved.

See also "Spiders".

Text reduction / Automatic Abstracting

(Oracle, Architext)

Conversion of the text to telegraphic style for speed reading. Creation of a short summary or abstract of a long document.

Application Programmer's Interface (API)

(Oracle)

A documented interface for programmers, with a library of functions.

Hardware and Software Requirements

(see specifics for each vendor)

Run on multiple platforms (e.g. Sun, Windows)? Require or not other software?

```
TRW-search      Not yet released
Verity          $10,000 per server plus $1800 for support.
WAIS            Free server and client software available (free-WAIS)
                Commercial version, $15,000 for the 1st server and $7,500
                for additional servers. One year of free support. $2500 for
                each additional year.
                Free trial of commercial version of 2 months.
```

7 Conclusion and future work

Chosing a search engine may be a rather complex and purpose driven decision. The feature glossary which is appended will help you making your own check list of features. We recommend that you contact the vendor to verify whether the features that you critically need are or not included. The next logical step is to download the most promising engine for a free trial period. If possible, build your entire application around this engine to test is thoroughly. In a second pass, you might want to reconsider other engines, in light of this first experience.

There are other search engines, such as the ones provided by OCLC and SIRSI which are not specifically designed form web browser clients, but are powerful tools that must be also considered for an electronic library which will have to organize massive volumes of data.

Before you purchase an engine, ask whether there is an AT&T rate which has been negotiated. This is the case, for instance, for Fulcrum.

In the future, it would be desirable to evaluate search engines with calibrated benchmarks. Please, send your engine evaluations and suggestions to improve and update this document to isabelle@research.att.com.

Acknowlegments

Many people contributed comments and useful suggestions and, in particular, I would like to thank John Denker, Fernando Pereira and David Lewis at Bell Labs, Rob van Meter and Jeff Pickard at AT&T GIS, Doug Monroe from QUEST Bob Waldtein from LINUS and all the people who answered my questions on the "hotair" mailing list. To subscribe to that list, send a message to: att.infosystems.www-request@hotair.att.com with 'subscribe' in the Subject line and the 1st line of the body.

A FEATURE GLOSSARY

Features that we may want to check. These were advertized in the Web pages of the various vendors. In parenthesis, the name of the vendors who claimed that feature (which does not mean that other vendors do not have it).

	Query-by-theme. Master index. API.
PLS	Specialist of text retrieval. Almost any feature, including distributed database model.
	Unclear whether it can couple with relational databases.
TRW-search	Manages other search engines and databases (not a search engine itself). Provides a unified interface.
	Virtual databases.
Verity	Topic agents. Query by example.
	Integrated with netscape server.
	Unclear whether it can couple with relational databases.
WAIS	Search engine incorporated into its own server. Web interfaces exist. NLS mixed with boolean queries. Relevance feedback.
	Server on UNIX, but clients anywhere.
	Many binary format filters available.
	Unclear whether it can couple with relational databases.

5 Years experience/partners/number users/famous users

Architext	Architext was founded in September of 1993 by six Stanford University graduates.
Fulcrum	Since 1983 indexing technology provider.
	Customers: Microsoft, Fujitsu, National Semiconductor, Florida Power and Banca di Roma. Licenced: Saros, Interleaf, Frame, Corel.
Oracle	12,500 software professionals in 93 countries, 3,000 partners and third party products supporting Oracle product in general (not necessarily WWW related).
PLS	12 year experience.
	America Online, Dialog, Dow Jones, DataTimes, Financial Times Information Services, Congressional Quarterly, Apple, Grolier Electronic Publishing and MacMillan New Media/Paramount
TRW-search	The Business Intelligence Systems branch of TRW was founded in 1994.
Verity	Founded in 1988. Used by 650 corp. and org.
	Partner with Netscape. Used in Adobe Acrobat and Lotus Notes.
WAIS	Project stated in 1989, WAIS inc. since 1991. Partner with Fulcrum.
	900 servers in 12 countries. 120,000 users in 28 countries (as of Oct 1994).

6 Price/free version/free trial

Fulcrum	$15,000 per server plus $2500 for support for the Surfboard product.
PLS	$5000 plus $750 for support for the PLWeb product.
	Free 45 day evaluation.

	thompson@verity.com
WAIS	Patty Johnson
	415 356 5481

A Fulcrum representative who has been dealing with AT&T a lot is:
Eric Treworgy
Sales Manager, North-East Region
Fulcrum Technologies, Inc.
67 South Bedford St.
Suite 400 West
Burlington, MA 01803
ph. 617-229-5824
fax. 617-272-0558
mobile: 617-281-1896
email: erict@fultech.com

3 AT&T users contacted

Fulcrum	AT&T GIS
	Rob van Meter
	513 847 4338
	Jeff Pickard
	513 847 4449
PLS	http://www.tollfree.att.net
Verity	Polly Yu
	908 615 4511
	ppy@mozart.att.com
WAIS	QUEST
	Steve Simon
	908 949 5156

4 Description/main competitive advantage claimed by the company

Architext	NLS, Boolean and SQL queries. Automatic abstracting. Distributed database model. Automatic subject grouping. Small index overhead.
Fulcrum	NLS, Boolean and SQL queries. Couples with relational databases. Multinode server searching.
Oracle	Couples with Oracle relational databases and their search engines and other Oracle tools. NLS and SQL queries.

1 Addresses

Architext	http://www.atext.com/ 2700 Garcia, Mountain View, CA 94043 415 934 3611 info@atext.com.
Fulcrum	http://www.fultech.com/Eng/corp/corphome.htm Fulcrum Technologies Inc., 785 Carling Ave. Ottawa, Ontario, Canada K1S 5H4 613 238 1761
Oracle	http://www.oracle.com/ 500 Oracle parkway, Redwood Shores, CA 94065 415 506 4514 INFOTEXT@us.oracle.com
PLS	http://www.pls.com/ Personal Library Software 2400 Research Blvd., Suite 350, Rockville, Maryland 20850 301 990 1155 info@pls.com
TRW-search	http://www.bis.trw.com/ TRW, BIS, 1329 Moffett Park Drive, Sunnyvale, CA 94089 800 767 8457 smart-info@bis.trw.com
Verity	http://www.verity.com/ Verity, Inc., 1550 Plymouth Street, Mountain View, CA 94043 415 960 7600 moreinfo@verity.com
WAIS	http://www.earn.net/gnrt/wais.html http://www.wais.com/ For WAIS customer support, contact support@wais.com WAIS Inc., 690 Fifth Street, San Francisco, CA 94107 415 356 5400 info@wais.com

2 Sales representatives contacted

Fulcrum	Christine Cowen 415 802 7054
PLS	Mary Hyman 301 208 12 49 mary@pls.com
Verity	John Thompson 415 960 7717

Survey of commercial search engines for WWW servers

Isabelle Guyon
AT&T Bell Laboratories, Holmdel, NJ 07733

September 28, 1995

Abstract

This report lists the most widely used commercial search engines for organizing World Wide Web sites on the internet. A server running one of these engines gives access to a number of document databases via simple queries in natural language. We include marketing information, pricing, contact names of sales representatives and AT&T users. We provide a feature glossary.

Keywords: search engine, commercial search engine, topic search, World Wide Web, HTML, information agents, security, query, database, electronic library, indexing, information retrieval.

Introduction

With the rapid growth of the internet, there is an increasing interest in making electronic libraries accessible through the World Wide Web. Even within the "firewall" of a corporation, it is possible to use the elegant HTTP protocol without alienating security.

There exist several freely available search engines suitable to organize Web pages (see e.g. WAIS (Thinking Machines Corp.), SWISH (EIT), Isite (CMDR), Glimpse (U Arizona), Aliweb (Nexor).

The Bell Labs library offers a search package called Slimmer which is used to organize LINUS and http://www.att.com. For more information on this engine, contact: Bob Waldstein, wald@library.att.com, 908 582-6171.

To organize larger Web sites, some business units of AT&T prefer licensing an "off-the-shelf" commercial engine, easy to install, supported and maintained. This report provides a set of pointers to information regarding the most widely used commercial search engines for WWW servers. Technical terms and abbreviations are appended.

We are aware that this survey is not exhaustive. New products appear almost dayly on the market. In particular, many universities are starting to commercialize academic search engines. Good examples include: "Inquery" of the University of Massachussets Amherst which is the engine used in Infoseek and the Library of Congress THOMAS system; "Pursuit" which is the search engine of "Lycos" the famous spider of Carnegie Mellon University.