



SearchKey Technology

WHITE PAPER

Redefining the Value of Information

with

Search Engine Technology

ASTAware SearchKey™ Technology White Paper

The Company

ASTAware Technologies INC. ("ASTAware") is a Canadian-based technology company, founded in 1988, and is publicly traded on the Vancouver Stock Exchange under the symbol AWA. ASTAware has developed an expertise in information management systems over the past twelve years. Our expertise in search, index and retrieval functionality is embodied within our two proprietary technologies: SearchKey™; and InNavigator™, positioning the organization at the leading edge of the information retrieval market. These technologies are deployed in a suite of products and solutions targeted to satisfy the diverse needs of the global marketplace.

The Company has a blue chip customer base located throughout the world on four continents that represents a varied cross-section of business and government. The sectors include Transportation, Telecommunications, Energy, HealthCare, Technology, Manufacturing, Publishing and Financial Institutions. Our customers include: Air Canada, Environment Canada, Hewlett Packard, IBM, Lockheed Martin, Monsanto Company, Omron Electronics, Ontario Hydro, O'Reilly & Associates, Siemens, Silicon Graphics, Sun Microsystems, Thompson Publishing, the Toronto Dominion Bank, Transport Canada, Veritas Software Inc., Visa International, Union Gas and more.

ASTAware Search Technology

ASTAware's product line offers a platform independent Java™-based Information Retrieval Technology for Enterprise Knowledge Management. The products are fully customizable, sophisticated, designed for flexible expansion of features and easy integration with other technologies.

ASTAware, using our Java-based technology, has developed high-speed search and retrieval software tools that assist businesses in organizing and navigating information in a web-based, e-commerce environment. Our technology is readily used by software developers to build search and retrieval capabilities into their own web-based applications and is easily integrated and interfaced with other technologies and products.

Indexing

In order for information to be searched, it must be indexed. ASTAware's SearchKey PRO Information Server indexes documents residing on the local file system, on the intranet, and on Web servers anywhere around the world, as well as database repositories. Since indexing is performed without duplication or modification of the document data, the documents remain in their original form and location. Many document formats can be indexed, and a URL to each document is stored in the index. Field indexing through metadata fields, as well as defining a paragraph or page as a single unit of information is supported. An unlimited number of documents can be inverted into a single index, and multiple indexes can be searched concurrently in response to a single search query or multiple queries, and seamlessly return a result list of documents with relevance ranking. This allows for a very efficient categorization of document groups, and search queries can be targeted to a single category or combination of categories.

Once an index has been created, and some of the indexed documents have been modified, the index can be updated through an incremental index update. This requires the re-indexing of the

modified documents and any new documents to be added to the collection, without rebuilding the entire index.

Through the use of Indirect Indexing, that is the indexing of a meta-document describing an entity that would otherwise be difficult or impossible to index (e.g. an image or sound file, records in a relational database, etc.), the indexing of the entity is accomplished indirectly.

Indirect indexing is also used to index databases, thereby allowing the simultaneous searching of legacy databases as well as collections of documents residing in the file system of any server on the network. Indirect indexing facilitates the indexing of a view of relational database tables, along with all other supported document types. Specific attributes from the database tables are selected and indexed as SearchKey fields. Thus, the search requests are performed at optimum speed on SearchKey indexes, regardless of where the information resides or originates from. When database tables change, the modified records can be re-indexed in SearchKey through an incremental index update.

Document Parsers

The indexing of specific document types is performed with ASTAware's document parsers. The document types supported include HTML, XML, Postscript, Acrobat PDF, Microsoft Word, Microsoft PowerPoint, Microsoft Excel, Digital Paper, and ASCII text.

The database administrator has the option of choosing which metadata fields to index as unique fields for searching where appropriate, in addition to main body text.

A general-purpose parser is available, whereby an "unknown" file extension is identified, and the software is instructed to treat those file types as if they are another. For example, files of type Java can be indexed as text.

Additional parsers can be added to support other document types as required.

ASTAware SearchKey PRO Information Server

The ASTAware SearchKey PRO Information Server is comprised of an Administration Server Service, a Search Server Service and an interactive GUI interface. The two services are responsible for building and managing indexes and for handling search requests coming from client machines. These services include a comprehensive protocol that is used to request actions and to deliver responses from the two servers. The search protocol is a public protocol that is supplied with the product when purchased. The administration protocol is a private protocol that is not generally distributed with the product, but is available publicly in the SearchKey API. Both services use a TCP/IP socket (usually 6010 and 6020 by default) to communicate with clients except in the case of the API which offers other communication mechanisms. This means that the client machines establish a private connection with the server to exchange information, and the services run independently of other processes on the server.

Administration Server Service Tasks

The Administration Server Service's main tasks include creating and editing indexes, setting up and maintaining the search service, and performing some server configuration and performance settings duties at the request of the administrator. Administrators communicate with the server through a GUI interface that is launched with the SearchKey PRO Information Server. It is also possible to communicate with the server through a remote installation of the SearchKey PRO Information Server. The GUI guides administrators through the required tasks such as setting the indexing parameters and activities of the server.

The primary function of the Administration Server Service is to invoke and manage the indexing process. This process begins with identifying the servers where the information is located (typically the web root directory on each server) and the web address or URL associated with the site. In order to index documents located on multiple servers, the administrator must have access to the hard drive of the server from the administrative client. Once these locations are identified, there are other parameters that must be set in conjunction with the building of the index. The first of these is naming your index. This enables you to easily keep track of multiple SearchKey indices and clone indices for updating your index while the search service is uninterrupted.

Selecting the documents for the build:

The selection of the documents to be included in the index is done by browsing the directory structure of the server (or servers) where the information is located and selecting individual files or entire folders to be indexed. The selection of documents on remote sites is done with the use of a built-in web crawler. These documents can be of the types as outlined in the section on document parsers.

Setting parameters:

The Administration Server Service has a built in scheduler, enabling unattended index building. The administrator can specify the time and date for an index build, even requesting that it happen every day, week, month at the specified time. The index can be activated immediately upon its completion, or can be activated manually by the administrator. Another parameter requiring attention is Stop Words. These are words that are extraneous to a search. For example the word "the" adds little to the search process and therefore, administrators generally choose to omit that from the index to shorten the index building process and increase efficiency. A default list of Stop Words is provided, and this can be changed as required.

The building of an index:

The Administration Server Service uses a simple file list for indexing. The file list is created automatically by the administration tool. The indexing process will accommodate for content from multiple servers, which can be indexed together. For example, the input file list may look like this:

```
http://www.abc.com/index.html  
http://www.abc.com/data/index.html  
http://www.abc.com/data/data1.html  
http://www.abc.com/data/test.txt  
http://www.server1.com/index.html  
http://www.electronic1.com/example.html
```

This list would be used by the Administration Server Service to access the files from the three different servers over the HTTP file protocol. The server would build an index of the documents on each of the servers listed without needing to download the contents. The indexing process involves opening the documents, reading the contents and building an index of all the keywords and numbers that are within. A major component of the indexing process is the parsing and extraction of terms from different file types and those defined by the Stop Word list. The resulting index is compact and efficient, as it is generally less than 10% the size of the original documents.

The Administration Server Service is also used to manage the search loads on the server. Up to 100 simultaneous search threads can be created on the server at one time and unlike CGI search applications, the threads are not destroyed, but can be reused.

Search Server Service Tasks

The Search Server Service responds to incoming search requests through a TCP/IP socket. The Search Server Service is capable of handling simultaneous requests and passing the results of the search to a results formatter. The results formatter passes the results back to the user (or client), or directly as a results stream. The search protocol is used to assist the Search Service in understanding the nature and type of query that the user has requested. This protocol allows end-users to use options that help to define a search query more clearly, for example to narrow the search to a specific Meta field, or a specific index.

Query Types:

Users can structure queries using Boolean operators, they can select how they would like to see the search results presented, whether they want to perform a Cascading search, or to accommodate for fielded searches where appropriate. For example, the protocol takes into consideration the Boolean operators (and, or, not, exact phrase, near, exclusive-or) that are typically sent in a query string e.g. "computer AND program NOT software" and treats the operators with the precision required. Queries can range from very simple to as complex as required. The results of a search can be further refined through a search within search function. Wildcard searching is also supported. For example, users can enter Java* as a query and their keyword query will include all words that begin with Java and have various extensions. Other forms of wildcarding (front- and middle-masking) will be added as required.

Exact phrase search, where an arbitrary phrase is matched word for word in documents, without any loss in search engine performance, is supported.

Boolean Search

Boolean search technique is a powerful server-side search solution that offers lightning-fast access to information on database and indexes. Boolean search is sophisticated enough to find even the most obscure information on large databases and web sites, yet flexible and rich-featured to accommodate the specialized needs of smaller sites. The cornerstone of the product's uniqueness is its dedication to "intelligent search" and to "search engineering" methods, which fine-tune the relevancy and precision of search results.

Boolean search techniques may be used to perform accurate searches without producing many irrelevant documents. When performing a Boolean search, SearchKey searches the computer database for the keywords that best describe the topic. The power of Boolean searching is based on combinations of keywords with connecting terms called operators. The three basic operators are the terms AND, OR, and NOT.

AND Operator

The operator AND narrows a search by combining terms and retrieves every document that contains both of the words specified.

OR Operator

The OR operator broadens or widens a search to include documents containing either keyword. The OR search is particularly useful when there are several common synonyms for a concept or variant spellings of a word.

NOT Operator

Combining search terms with the NOT operator narrows a search by excluding unwanted terms.

Complex search

SearchKey has complex search capability using more than one operator, where search terms can be nested. Search terms and operators included in parentheses will be searched for first, then terms and operators outside the parentheses.

A search for: (ADD OR attention deficit disorder) AND college students will search for documents containing either the acronym ADD or the keywords attention deficit disorder, then narrow the search results only to those documents which also contain the words college students.

Wildcard

SearchKey offers a facility to search the database for documents where minimal knowledge of document details is available. Use the wild card search from within any search form. Wildcards can be used in one or more fields.

Wildcard searches are activated by inserting a "*" (asterisk) sign.

Phrase Search

The exact phrase is a handy tool for finding specific documents. An exact phrase enables you to search on documents that have a specific sequence of words such as 'mutual funds' or 'investment broker'. To type an exact phrase into the search engine you only need to enclose the phrase you want searched in quotes. The search engine will only return the documents that contain the full phrase and not the individual words. This technique is useful for narrowing down documents where there are many common words that appear frequently on most documents. Thus searching with a single keyword will return an enormous amount of documents. With phrase search this result can be narrowed to minimum results.

Search Within Paragraph

The paragraph searching is similar to the Boolean Search with AND operator but more powerful in a sense that the search engine returns only documents that contains these words in the same paragraph. Paragraph searching enables users to search for combination of words within sentences or paragraphs.

Search Within Search

SearchKey has added value to its current search engine by allowing users to search from within search results. As databases grow the necessity for the user to continually narrow a search to the item that they truly want will grow also. Upon the first search users will have the option to perform a search within the current results. Searching within the current results will take the first search and add it to the new search and so on.

Thesaurus Lookup

A thesaurus lookup capability to find words or phrases related to a keyword of interest is available in SearchKey.

Targeted Search Interfaces:

Site administrators can design search query interfaces using HTML forms providing text fields and other standard GUI elements making choices available to the user to form an appropriate search query. The submission of the search form to the Search Server will trigger a search and the results formatting process.

Cascading Searches:

SearchKey PRO Information Server offers a very flexible and powerful enterprise search system allowing for independent indices, each located on a separate server anywhere around the world, to be tied together into a single knowledge base. The Search Server passes a request to other registered servers that have been identified by the administrator. The secondary servers execute a local query on their indices and send search result responses back to the first host. This method of simultaneous, distributed searching is called Cascading Searches. The administrator identifies the location of the alternate indices when setting up the parameters of a particular index.

SearchKey PRO can extend the search capabilities of ASTAware's SearchDisc product, a search engine for CD/DVD's, by allowing SearchDisc to cascade search requests to alternate servers running the SearchKey PRO Information Server.

Results Formatting and Presentation:

When a search request is received by the search service, an index lookup is performed and the results obtained from the index are passed through a formatting process. This creates a dynamic HTML document set that is controlled by a standard or custom template. The formatted dynamic HTML pages are returned to the client and viewed with a normal browser. Highlighting of search terms and phrases is performed on documents in HTML format.

SearchKey PRO offers two standard templates. Simple Search Results offers a list of found documents (in groups, e.g. 1-10, 11-20, etc.), comprised of document title, relevance of each document indicated by the ranking of frequency counts of the found keywords, and document size. Expanded Search Results offers the document title, the first few lines of the document's content, the document URL, relevance of each document, and document size. Administrators can customize this presentation by modifying the HTML template using the SearchKey PRO Template Editor.

Administrators can also create fully custom templates with the option of displaying any combination of the following parameters: document title, document URL, the first few lines of the document's content, document relevance indicated by the frequency counts of found keywords, document size, content of specific Meta Tags where appropriate, current page number, link to next page, link to previous page, links to every page in the set, total number of pages, number of hits per page, total number of hits, and maximum number of hits allowed.

Scalability

ASTAware's search solutions are highly scalable. The architecture of SearchKey PRO is designed to handle the indexing of millions of documents, and searching of the collection concurrently by many users. The number of documents that can be searched through a collection of SearchKey indices is virtually limitless with the multiple index search module that allows multiple users to send multiple queries to multiple indices and generates a single list of results for each user. Through the powerful Cascading Searches functionality, allowing the simultaneous, distributed searching of indices located on servers around the world, SearchKey PRO is a powerful enterprise search system tying communities of servers into one single knowledge base.

ASTAware's SearchKey technology comes with a rich set of search functions that can be accessed through a search protocol that is easy to use and precise, and also extensible to allow the implementation of new functionality as required.