Searching Structured and Domino Databases
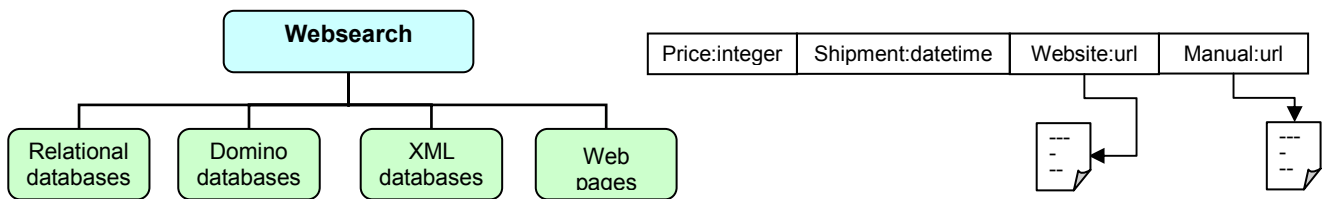
Search@

Search is the core of our business … and <u>yours</u> too!

Searching Structured Records and Beyond

A structured record consists of a number of fields organized according to a fixed schema. Data in a relational database are structured records in that a table defines a fixed set of fields (columns) that every row in it must follow. Furthermore, each field has a fixed, atomic data type and is indexed as a single value. For example, a text field holding an entire document is treated as a long text string and indexed as such. Therefore, creating an index on a long text field in a relational database is useless if you are looking for individual words within the text field.

On the other hand, search engines for free text index a long text string as individual words. They are very efficient in searching word combinations within a long text string. However, they are very limited in searching structured data. Most search engines can search only simple fields such as title and last modification date with no support on inter-field Boolean operators.

Websearch's field search and full-text search capabilities support both worlds and go beyond. It provides a uniform set of search functions across a wide range of structured and semi-structured databases, shown in the chart below. Through JDBC, an industrial standard for accessing databases, it can index and search data from any combination of tables stored in relational databases. It can index and search fields and attachments of Domino databases. It searches XML records as well as meta-tags in web pages.
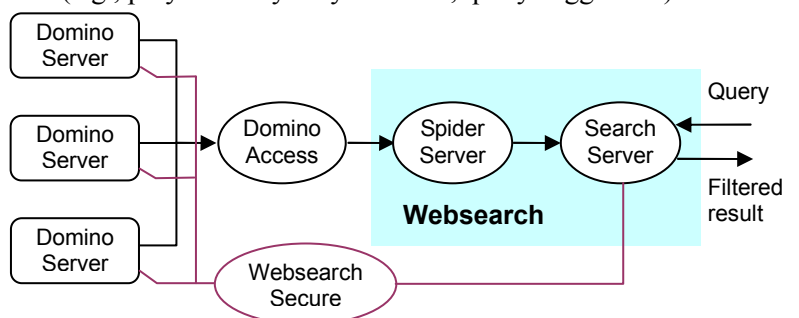

Websearch allows full-text and field search to be used together. It indexes text fields by breaking then down into words; it follows and indexes files pointed at by URLs, thus enabling search on fields and attachments at the same time.

Highlights of Field Search Features

- Index embedded documents referenced by URLs inside databases
- Fields can be grouped together for indexing and searching as a single field
- A field can be selected to serve as the title, summary, last modification date or document length of a document when it is displayed
- Different data types are supported:
  - ☐ int: integer values supporting equality, relative and range comparisons
  - ☐ datetime: data/time values supporting equality, relative and range comparisons
  - ☐ text: text strings supporting text comparisons without stopword removal and stemming
  - ☐ webtext: same as text but with stopword removal and stemming
- Boolean operators can be applied between fields and within fields
- Results can be sorted in ascending or descending order on multiple fields (e.g. sort by score in descending order; if the scores are the same, sort by the name field in ascending order.)
- Results can be displayed with keyword highlighting
- Fully integrated with Suntek's search modules (e.g., pinyin and synonym search, query suggestion)

Websearch for Domino

A Domino document consists of a number of fields and optionally a number of attachments. To search a Domino database, you need a search engine that can search document fields and the full text of the attachments. Websearch for Domino is an

add-on module that extends Websearch's powerful field search and full-text search to Domino databases. In the diagram below, Websearch interacts with Domino servers through a common access module, which retrieves documents from Domino databases via CORBA/IIOP and passes them to Websearch for indexing. Users can search Domino documents in the same way as other documents hosted on other platforms. Since in most corporate environment user access to documents is restricted, Websearch Secure can be incorporated to remove unauthorized documents from the result page. Together with Websearch Secure, Websearch for Domino is an integrated search solution for their Domino databases in corporations.

Highlights of Websearch for Domino

- A single Websearch for Domino can index and search multiple domino servers. It is ideal for creating a single access point to all of your Domino databases across your entire organization.
- Domino databases to be indexed can be conveniently specified with regular expressions. For example, the regular expression "`\/public\/*\.nsf`" specifies that all Notes files under the "`…/public/`" directory are to be indexed.
- Suntek's powerful field search and full-text search are supported. In particular, to cater for the diversity of a corporate environment, different document fields can be grouped together for searching. For example, different Domino databases may use different field names such as "docTitle", "chiTitle" and "engTitle", etc., for the document title. User can define a "searchTitle" field in the search engine to search different titles in different documents.
- Websearch interacts with Domino servers through Cobra/IIOP; no change is needed on Domino servers.
- Integrate with Domino's access control system to support authenticated search on Websearch.

Websearch for Domino Security Features

- Supports database and document level access controls in accordance with the policies in Domino Server.
- Minimized filtering time with caching and optimized filtering algorithm.
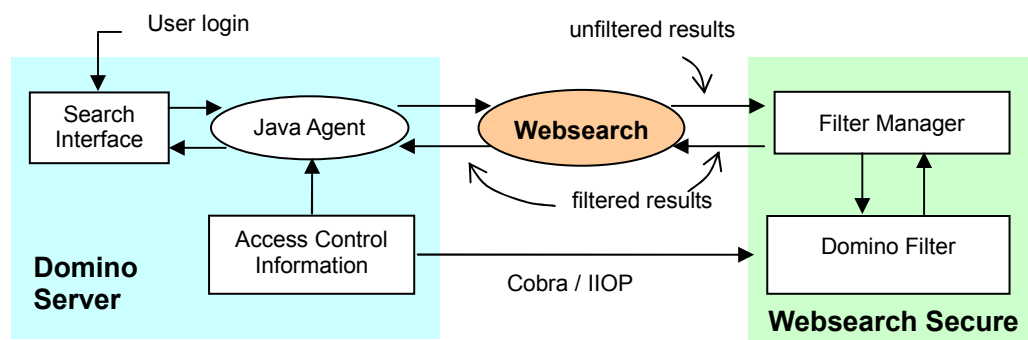
The Search Interface can be accessed through a web browser or a Lotus Notes client. Either way, users have to be authenticated by the Domino server. Once successfully authenticated, users can submit queries from the search page. The Java Agent will transmit the search queries together with the user names and the associated user groups to Websearch for searching and filtering.

Websearch will conduct the search as usual and pass the unfiltered search result, together with user name and user groups, to Websearch Secure, which in turn dispatches all the information to the Domino Filter for security filtering. The Domino Filter obtains access control information from the Domino Server and filters the results based on the user name, the user groups and the access control information. To enhance filter performance and avoid overloading the Domino Server, The Domino Filter caches a subset of the access control information in the cache and periodically synchronizes the cache with the Domino Server.

The Domino Filter interacts with Domino Server through Cobra/IIOP. Therefore, the Domino Server can reside anywhere on the network and does not require any external software or modifications.

Company Profile

Suntek is a software company specializing in search and content-management solutions for Asian languages, especially Chinese. Owning all the source codes, Suntek can provide high-quality, fast turnaround support for and customisation of its products. Major customers include *The Hong Kong Police Force*, *The Hong Kong Government Information Center*, *Independent Commission Against Corruption* (*ICAC*), *Hong Kong Securities and Futures Commission* (*SFC*), *TVB.COM*, *Orange Hong Kong*, *The University of Hong Kong*, *Hong Kong University of Science and Technology*, *City University of Hong Kong* and *Hong Kong Institute of Education*.