SPOT: TRW'S MULTI-LINGUAL TEXT SEARCH TOOL

Peggy Otsubo

TRW Systems Development Division
R2/2162
One Space Park
Redondo Beach, CA 90278
peggy@wilbur.coyote.trw.com

ABSTRACT

TRW has developed a text search tool that allows users to enter a query in foreign languages and retrieve documents that match the query. A single query can contain words and phrases in a mix of different languages, with the foreign-language terms entered using the native script. The browser also displays the original document in its native script. Key terms in the browser display are highlighted. The interface is targeted for the non-native speaker and includes a variety of tools to help formulate foreign-language queries. Spot has been designed to interface to multiple search engines through an object-oriented search engine abstraction layer. It currently supports Paracel's Fast Data Finder search engine, with support for Excalibur's RetrievalWare currently being developed.

1.0. INTRODUCTION

1.1. Design Objectives

TRW has developed a text search tool that allows users to enter a query in a number of languages and retrieve documents that match the query. This text search tool is called Spot. The following subsections describe the design objectives and goals of Spot.

1.1.1. Support multiple search engines

Our government users currently use a variety of tools for different purposes. For example, an archival database is only available through a legacy text search system that performs its searches very quickly, but lacks a great deal in search functionality. Other users use Paracel's Fast Data Finder search engine due to its powerful search capabilities and are only able to access its power through the FDF search tool user interface.

One of our design objectives was to handle multiple search engines within the same user interface tool. This provides users with a single user interface tool to learn, while providing them with a choice of search engines. Users might choose to perform a natural language query using the Excalibur/ConQuest search engine's concept query and switch to the Fast Data Finder to search Chinese text.

We also aimed to provide the users with the full functionality of each of the search engines. This approach necessitates a more generic approach to many functions to ensure that the same user interface can be tailored to differing search engine technologies.

1.1.2. Support multi-lingual data

Internationalized support is fairly easy to obtain commercially for a number of commonly-supported languages. The commercial products for internationalization are designed to support the marketing of a tool in a specific set of foreign countries, where the menus, buttons, error messages, and text all need to be displayed in the appropriate foreign language. For example, if a specific product needs to be marketed to the Japanese, it might be running under Sun's Japanese Language Environment, with JLE providing support for entering and displaying Japanese text.

Multi-lingual support, however, is very difficult to obtain commercially. Our user community consists of native-English speakers, who want the menus and buttons to appear in English, but require support for viewing foreign-language documents in their native scripts, as well as entering foreign-language query terms in their native

scripts. For this functionality, internationalized support is inadequate.

1.1.3. Support query generation tools

Users who are not native speakers of the foreign language in which they are submitting a query would like tools to assist in building queries. For example, we located a large Japanese-to-English thesaurus that was available in electronic form. It would be very useful for native-English speakers to look up relevant words in the Japanese thesaurus for assistance in building their queries.

In addition, words that are of a foreign origin are often transliterated in a number of different ways. For example, the name "Kadafi" is often spelled "Khadafi" or "Gadaffi". Query generation tools that allow users to enter "Kadafi" and find the other possible spellings are designed into Spot.

1.2. Maximize performance

Spot was designed to be the user interface for a large archival database of hundreds of gigabytes of data. It needs to provide hundreds of users with access to this database.

An archival database using the Fast Data Finder was implemented using Paracel's Batch Search Server (BSS) product. Spot currently interfaces to this FDF archival database. Development is currently proceeding to interface Spot to an Excalibur/ConQuest archival database.

Our objective in developing functionality, including multi-lingual query generation tools and query functionality, has emphasized solutions that work very quickly, usually by exploiting the features of a specific search engine.

Speed and throughput of searches through the FDF hardware search engine was measured using a commercial FDF-3 system.A single FDF-3 produced a search rate of around 3.5 MB/s, which could be obtained while searching 20 to 40 average queries simultaneously. A system of multiple FDFs can linearly expand the search rate.

1.3. User Interface Highlights

Some of the highlights of our current user interface system include the following:

- Multi-lingual query entry
- Multiple languages in a single query
- Queries can be saved, loaded, edited, printed
- Customizable fill-in-the-boxes query form
- Query generation tools
- Highlights query terms when browsing search results
- Display of search results in native script
- Copy-and-paste from Browser into a Query
- Search using Paracel's Fast Data Finder
- Search using Excalibur/ConQuest's RetrievalWare

2.0. DESCRIPTION OF THE SYSTEM

Spot is a TRW-built graphical user interface tool that supports query entry and browsing of search results. Some of the goals of Spot include the following: support multiple search engines by allowing users to select the search engine they desire, support for multiple languages within the same tool without requiring different versions of Spot or different operating system support, query generation tools to assist non-native speakers in creating foreign-language queries, capability of browsing foreign-language text using their native script, and support for utilities to aid analysts.

A full set of features to browse, save, recall, print, and manipulate the results of searches are provided. There are also features to select the databases to search and to create and load new databases. Each screen has a pulldown Help menu, which brings up a description of the available functions.

2.1. Support for Multiple Search Engines

We allow users to select the desired search engine by merely selecting the desired search engine from a pulldown menu. The list of query forms, query functionality, and databases available for searching is dependent on the search engine that is selected.

Spot consists of a Search Server Interface layer that uses an object-oriented approach to abstract the details of each search engine from the user

interface portion of Spot. Figure 1 illustrates how the Search Server Interface and the search engines fit into the architecture of Spot.

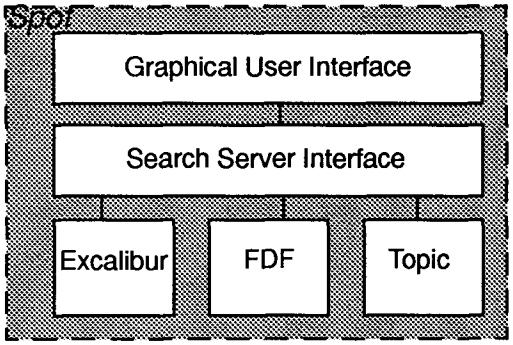


Figure 1: Multiple search engines in Spot

The Search Server Interface consists of a package of code routines that act as a broker for the search engine functions for Spot. These functions include initialization of the search engine, retrieving the list of databases that are searchable, retrieving a list of allowable search engine options, performing the search, and retrieving results.

A single software package contains code that directly interfaces to a specific search engine. For example, the SeFDF package contains all references to the FDF's API function calls.

2.2. Multi-lingual Query Entry

Since our government user community requires support of multiple languages in the same query, we can not rely on operating system support, such as the features provided by the Japanese Language Environment for Sun Workstations. To provide multi-lingual query entry support, we have integrated the New Mexico State University's multi-lingual text widget, called MUTT, to provide users with a full multi-lingual query entry capability. The latest release of MUTT, version 2.0, supports Unicode as its underlying character set. There is also built-in support for conversions between various encoding sets for a particular language. Spot uses this feature to allow users to expand the character sets across which to search.

The languages that are currently supported by NMSU's MUTT include: Japanese, Chinese, Korean, Latin, Serbo-Croat, Lao, Thai, Arabic, Armenian, Russian, Georgian, Hebrew, Irish, Portugese, Rumanian, Czech, French, Spanish, Vietnamese, Ethiopic, Latvian, Greek, Turkish, Icelandic, Italian, and Dutch.

Figure 2 illustrates Spot's Main Window. The portion of the window below the fill-in-the-boxes Query Editor is the multi-lingual text area. Users can select the language and the entry method using menus. The foreign-language text is entered in the text area just below the Language menu.

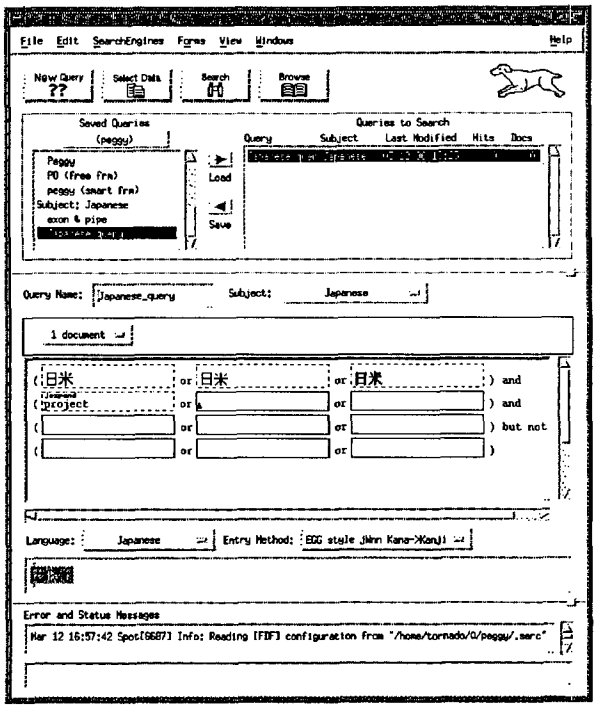


Figure 2: Fill-in-the-form multi-lingual query

The basic query formulation paradigm in the above figure is extended boolean search. The user is provided with an expandable boolean form in which to enter query terms. Foreign-language terms are typed into the multi-lingual text widget are in the lower portion of the figure. These terms are then entered into the appropriate box in the query. Terms from the browser window may be copied and pasted into the query window to help refine a query. The user may select from a menu of choices to require query terms to be within a specified proximity window.

Query formulation paradigms other than the extended boolean search demonstrated above can be supported by designing a custom form. An external file is used to specify the appearance and functionality of a form. For example, Spot also supports a term-weighting form that allows users to type in a term (a word or phrase) and assign a specific weight to the term. Results of this term-weighted search can then be ranked in order of score, presenting users with the most-relevant score first.

2.3. Query Generation Tools

One of our goals during the design of Spot was to provide support for non-native speakers, which will believe make up the bulk of our prospective government users. We implemented three types of query generation aids.

First, the user may select from a menu of predefined subqueries. These subqueries can be "canned" queries that are set up by the system administrator or an advanced user. Alternatively, the subqueries can be developed by the user himself and saved as a subquery for use in other queries. These subqueries include any valid query expression and provide a shorthand method of referring to a previously-defined entity. For example, the subquery 'Sony" may be used to search for references to the Sony Corporation and its products.

Second, Spot allows users to select appropriate terms for a search from a thesaurus. One of the thesauri that we have integrated into Spot includes a 28,000-entry English-to-Japanese thesaurus. Users select from the thesaurus in the following manner: the user enters an English word or phrase, selects an appropriate thesaurus, selects any combination of lines from a popup window that displays lines from the thesaurus that contain the user-entered word or phrase, and the selected terms are included in the query. The thesaurus can be easily extended and other thesauri may be added at customer sites.

Third, Spot includes support for expansions to modify user-entered search terms. The most powerful use of expansions is to handle transliterations for various foreign-language words and proper nouns. Our English-to-Japanese transliteration scheme is an example of an expansion.

Here's a description of the "Japanese-katakana" expansion: Foreign loan words and proper nouns are represented in the Japanese katakana phonetic alphabet, based on its pronunciation. Particular difficulties arise when the foreign words contain sounds or patterns of sounds that are not defined in the Japanese language. In these cases, there are a number of different ways the foreign loan word or proper noun might be expressed in Katakana. While the Japanese "spellings" for common foreign loan words or Western public figures tend to become quickly standardized (and thus could be included in the thesaurus), company names, new product names, and non-public figures are not likely to be represented in consistent Katakana across sources.

Our transliteration algorithm maps an English word to its most likely Katakana possibilities. The basic idea is to break the word apart phonetically and then substitute as many of the possible ways the sounds might be heard by a Japanese speaking person as alternatives. Figure 3 shows three simple examples[1].

```
ronald reagan  => [ロ|ロー|ル] [ナ|ネー]ル[ド|ッド] [リ|リー|レ|レー] [ガ|ゲー]ン
bill clinton   => [ビ|バ|バイ|ブ]ル ク[リ|ラ|ライ|ル]ントン
mary brown     => [マ|メー]リー ブ[ラウ|ロウ|ロー]ン
```

Figure 3: Some English to Japanese transliterati

This particular expansion exploits the character-level expansion feature of the FDF to expand the list of possible matches. We reviewed the performance of this algorithm on a sample list of 150 English last names and tallied that the program was picking up the academically correct variation 80-90% of the time.

2.4. Multi-lingual Browsing

The Browser supports a full set of features including support for displaying the documents in their native script, highlighting the query terms, scrolling through the hits, saving documents or search results, printing the document or selected portions of the document with or without the highlighted portions in the native script, mailing selected portions of the text, copying portions of the text into other applications as well as the Query Editor portion of Spot, and performing external processing on selected portions of the text.

Figure 4 illustrates a Japanese-language document that was retrieved as a result of a search. Notice the highlighted areas that identify query terms found in the document. This Browser is easy to use, with arrow turn-signals to scroll through the list of hits, documents, and terms. Users can either use the scrollbar or the Page up-down arrows to scroll through a single document. There are also hot-keys to quickly scroll through the list of hits or documents, as well as

---

[1] The notation "a [b | c] d" means an "a" followed by a "b" or "c" followed by a "d". Thus the user entered term of 'Reagan' will match on any of

リガン or リーゲーン or レガン or レーガン

type-in fields for moving to a specific hit directly or a slider for moving quickly to another hit.

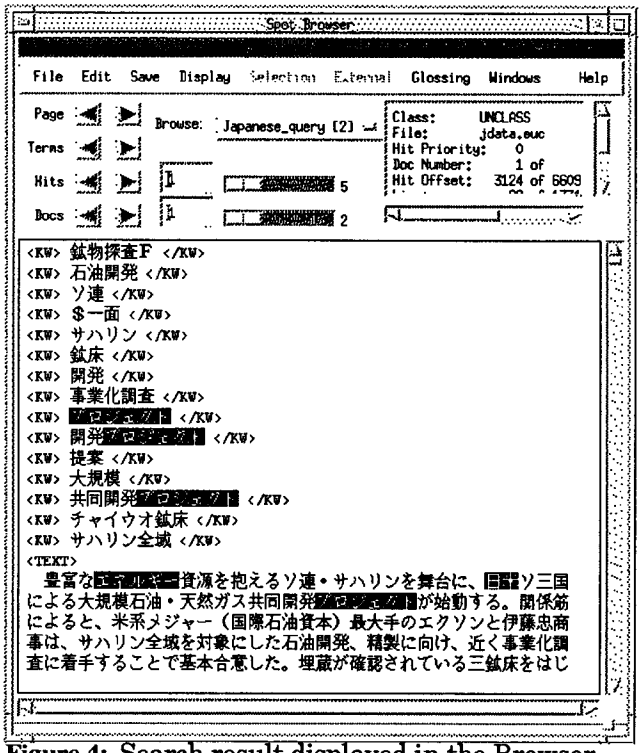


Figure 4: Search result displayed in the Browser

Users may view the original text in another "language" by switching the "Display" menu to another language. To view selected portions of text in another language, the user can highlight text and select another language in the "Selection" menu. This allows users to view a single document which contains multiple languages.

Another powerful feature of the Browser is the configurable interface to "External" processes. The user can highlight a portion of text, select an external process from the "External" pulldown menu, and view the results of applying an external process to the selected text appear in a popup window. This can be used, for example, to view a hex dump of the selected text or to view the directory listing of a phone number that was highlighted in the text. An "External" process can be easily added without modifying any of the Spot code.

2.5. Database Selection

Users select the database to search from the Database Window, which is illustrated in Figure 5. The data that is organized by date are shown in the upper portion of the window; the data that is not affiliated with a date is shown in the list in the lower portion of the window. The date-oriented data is shown in a spreadsheet, with an "X" in boxes with data. Users can zoom in to view the data by month or by day or they can zoom out to view the data by year. To select data for a particular day, they simply click on the box.

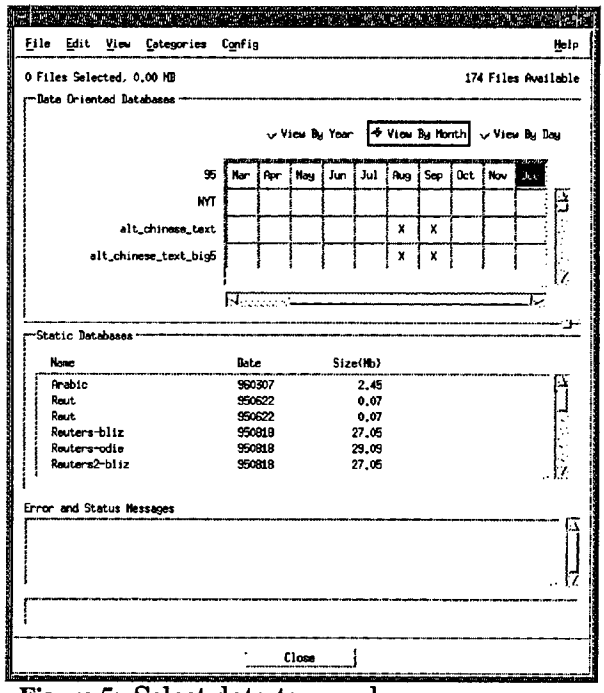


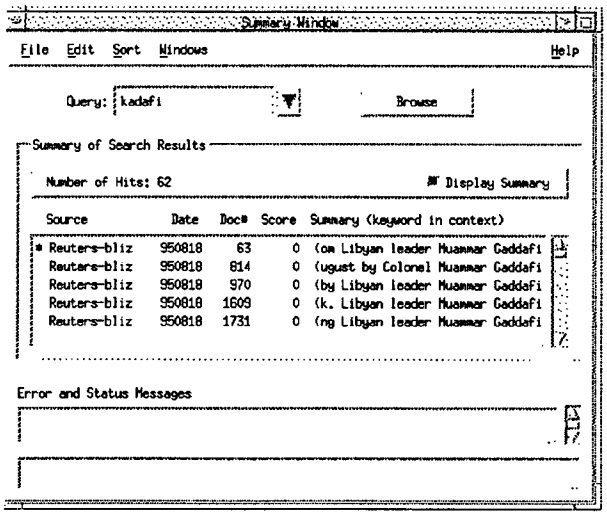
Figure 5: Select data to search

Most of our government users deal with timely data that is very strongly affiliated with a particular date. The spreadsheet portion of the Database Window has been enthusiastically embraced as the appropriate method for selecting data to search.

2.6. Summary Window

Users may also view a summary of their search results in the Summary Window, shown in Figure 6. Each hit result is described in a line, with information such as the date, document number, score, and keyword in context. Users highlight the hits that they would like to browse and push the Browse button to view the results. Hits that have been read are indicated with an asterisk in the first column.

The Summary Window is very useful for gaining a quick overview of the results. It is also useful for iteratively viewing the data by selecting the results to browse.



igure 6: Summary of search result hits

3.0. PAST EXPERIENCES

TRW has produced two other multi-lingual text search products that are predecessors of Spot. The experiences gained on these products were folded into Spot.

Our first experience handling foreign-language text search was the Japanese Fast Data Finder (JFDF) prototype that was developed as part of the Tipster program. JFDF used the Fast Data Finder search engine to search Japanese language data and used Sun's Japanese Language Environment (JLE) Operating System support for handling the Japanese query entry and browsing. This prototype was very successful with the analysts that dealt with Japanese data. However, expanding this capability to other languages was difficult, since the language capabilities depended on operating system support.

Our second experience with foreign-language text search involved modification of the FDF-based text search tool, called XATI, to handle Chinese. Our goal here was to support Chinese without using a custom Chinese-language operating system.

First of all, we modified the XATI browser to handle a number of different languages, including Chinese, Japanese, Korean, and Cyrillic. We built in support to handle 16-bit languages (Japanese, Chinese, Korean), as well as extended ASCII and other 8-bit languages. Modifying the Browser to handle these different languages was fairly straightforward, with much of the work dealing with collecting, building, and modifying fonts.

Then, we added support in the query editor to allow entry of Chinese. We built in support for the Pinyin entry method, using the public domain cxterm as our model. There are several encoding formats for Chinese. A configuration variable indicates which of the two major encoding formats to use (Big5 or GB).

The Chinese version of XATI worked successfully without direct operating system support for language-entry. We found that users wanted support for multiple encoding formats (i.e. both GB and Big5), as well as simultaneous support for additional languages. These user requirements were designed into our current multi-lingual Spot.

4.0. KEY INNOVATIONS

We have developed a multi-lingual text search tool that is being enthusiastically embraced by users. Some of our key innovations include:

- Search and retrieval of multi-lingual data, using queries specifying search terms in different languages and encoding sets.
- Display of search results in native scripts including Japanese, Chinese, Korean, Arabic, Cyrillic, Thai, and Vietnamese.
- Multi-lingual query entry using NMSU's multi-lingual text widget (MUTT).
- Multiple languages in a single query.
- Multiple encoding sets in a single query.
- Query generation tools to help non-native speakers build queries in different languages.
- Allow users to perform external processes on portions of browsed text.
- Fill-in-the-box, customizable query entry forms.
- Easy-to-use date-oriented database selection screen.
- Allow users to select their desired search engine.