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. 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.