DB2 Text Extender for your e-business

Search and Mining Functionality

DB2 Text Extender provides a set of search functions that meet all customer requirements in a vast majority of cases. These different functional characteristics are provided through different index types selectable based on the application requirements:

- A precise index for fast case-sensitive searches. No dictionaries are used
- A linguistic index for language specific processing using dictionaries, normalization and morphological processing
- A Ngram index with fuzzy (case-insensitive) search capabilities

Text Extender offers a wide variety of search functions dependent on the selected index type, namely:

For all index types:

- Section search. This limits your search to a specified section of structured documents (user-defined, HTML and XML).
- Boolean search. This allows for conjunction, disjunction, and exclusion of search terms. Individual search terms may be single words or phrases. Information needs can be specified very accurately.
- Proximity search. This allows you to specify that the search terms must occur in the same paragraph, or in the same sentence.
- Wild card search. Front-, middle- and end-masking can be applied using wild cards for single characters or strings of characters.
- Thesaurus expansion. This expands a search term to include new terms related to the search term. For example, search for "database" and also find documents that contain "repository" and "DB2". A small sample thesaurus is provided, but "no ready to use" thesaurus is supplied. With the thesaurus compiler that is part of Text Extender, domain and application specific thesauri can be defined and compiled for use in a search application.

For linguistic indexes:

- Base form reduction or stemming. This increases the recall of a search by expanding the search terms using a dictionary.
- Synonym search. These are language specific and extracted from provided dictionaries.

For linguistic and precise indexes:

- Phonetic search. This increases the recall of a search by expanding the search terms to include similar sounding terms. Certain characters in the search term are replaced with wild cards. For example, search for "gose" and find "goose".
- Free-text search. This is based on the probabilistic retrieval model and estimates the probability that a document is relevant given a query. A phrase or a sentence describes in natural language the subject to be seached for.
- Hybrid search. This is a combination of Boolean search and free-text search.

For Ngram indexes:

- Fuzzy search. This searches for words that are spelled in a similar way to the search term. Fuzzy search can be used to find names that have been incorrectly entered into a table or if the correct spelling is not known. For example, a search for "Andrew" can find "Andrews", "Andraw" and "Andru".

As the text search functions are SQL language extensions (user-defined functions), it is easy to combine full-text search with both parametric or multimedia (for example image) searches.
Text Extender provides parsers and filters for all well-known document formats and also automatically synchronizes the text index and DB2 contents.
Text Extender is closely integrated into DB2 access and control concepts, and also supports the DB2 EEE (Extended Enterprise Edition) environments.
It supports a modelling feature that recognizes the internal structure of a document and allows searches to be restricted to sections of the document.
You can index data stored either in DB2 tables or on files referenced using the DB2 Datalink Manager which also supports indexing and search on character data types, user-defined large objects and external files.
Linguistic indexing and search for 22 languages including English, German, French and Japanese. Base indexing and support for 37 languages.
Full-text index update can either be Incremental and asynchronous.

An example

A bookstore wants to build an e-commerce application for searching and ordering books. There are 200,000 records that contain author, title and subject information. The number of hits/day is expected to be between 100,000 and 200,000, that is, about 1 query/sec.

Because the author, title, and subject are multilingual, national-language character normalization is required. To support the new German spelling rules, a thesaurus is used. Increased recall is important because the bookstore wants to offer customers everything related to their search. Search performance is expected to be in the 1-4 second range.

DB2 Text Extender supplies the required linguistic search capabilities, as well as the desired performance.

Text document formats supported

- HTML, XML, RTF, MSWORD, WP5, your own format for structured documents

Platforms supported

- AIX, Solaris, HP-UX, Windows NT and Windows 2000, OS/2, z/OS and OS/390, and OS/400