DB2 UDB Text Information Extender V7.2
DB2 Text Information Extender for your e-business

Search and Mining Functionality

DB2 Text Information Extender V7.2 provides a set of search functions that meet all customer requirements in a vast majority of cases.

DB2 Text Information Extender V7.2 offers a wide variety of search functions, namely:

- Section search. This limits your search to a specified section of structured documents (GPP, 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 Information Extender, domain and application specific thesauri can be defined and compiled for use in a search application.
- Free-text search. A phrase or a sentence describes in natural language the subject to be seached for.
- 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 built-in SQL language extensions, it is easy to combine full-text search with both parametric or multimedia, such as image searches.

DB2 Text Information Extender V7.2 provides transformation function parsers for HTML, XML and GPP document formats. A user exit allows format converters to plug in and map the unsupported format to ASCII.

DB2 Text Information Extender V7.2 could automatically synchronize the text index and DB2 contents.

DB2 Text Information Extender V7.2 is closely integrated into DB2 Optimizer, which provides a very good performance when using several predicates.

It supports a modelling feature for 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.

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, year of publication and subject information. The number of hits/day is expected to be between 100,000 and 200,000, that is, about 1 query/sec.

Customers are able to search on a combination of metadata, such as author and year of publication and text data, like subject information and title using a variety of search mechanisms, such as fuzzy which allows searching on incorrect spellings. A thesaurus can also be used to include words with similar meanings in the search. 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 Information Extender V7.2 supplies the desired performance.

Text document formats supported

Ÿ HTML, XML, GPP, and plain ASCII.

Platforms supported

Ÿ AIX, Solaris, Windows NT, and Windows 2000.