Automatic indexing

Organising files into a directory hierarchy allows users to browse as well as searching. It is possible to generate hierarchies automatically, which could be useful if the number of files is large. For example, the results of a search could be indexed automatically and displayed in a hierarchy. Automatic indexing can also be applied to an existing directory structure to find alternative ways of organising the information.

The automatic indexing process has four stages:

1. **Phrase extraction:** relevant words and phrases are extracted from files and/or metadata. A list of phrases is made for each file, and a list of files is made for each phrase. Phrases that consist entirely of 'stop words' are not included. For example, the file "John Coltrane - The Definitive/09 - Acknowledgment.mp3" would be associated with the phrases 'john', 'coltrane', 'definitive', 'acknowledgment', 'mp3', 'john coltrane' and 'the definitive', but not 'the' or '09'.
2. **Sorting:** phrases are sorted by the number of matching files.
3. **Selection:** directories are created by selecting phrases in descending order of the number of matching files, until every file belongs to at least one directory or there are only a few remaining files, which are placed in the current directory.
4. **Recursion:** the sorting and selection stages are repeated for the contents of each directory, creating subdirectories.

m.rogers@cs.ucl.ac.uk
Last modified 2007/02/28