The Indexing of Technical Documents: the Use of a Knowledge Representation Model

The automatic indexing of technical documents sets specific problems. We are particularly interested in large, highly structured documents written in French and used by expert users in operative situations, but the methodology we propose could be adapted to other types of languages.

Information retrieval in technical documents presents specific characteristics (Paganelli, 1997). We have previously shown that expert users have specific needs; they search for information either about an object description, or about a procedural description. In addition, technical texts present linguistic particularities; they give information about objects and their characteristics, and about the actions made on or by these objects.

Our work takes place within the CRISTAL research laboratory which considers that the indexing of full-text documents should be realized by an automated system based on a natural language processing system. The CRISTAL system (Rouault, 1987) is based on a linguistic analysis which is concerned with retrieving morphological and syntactical structures.

We are proposing to complete this automatic indexing method by using a model of knowledge representation. In the natural language discourse, the manipulated objects are complex, and the reasoning is very often inferential (Berrendonner et al., 1992). The model proposes to give a structured representation of the objects of discourse and proposes reasoning methods that are either action schemata (Gallo et al., 1996) or state schemata. The action schemata fit with the procedural description and the state schemata fit with the expression of properties.

The indexing method we propose is made up of two stages:
- a morphosyntactical analysis identifies the noun phrases which represent the objects of discourse,
- an analysis based on the knowledge model gives a detailed representation of each object extracted.

It also needs a semantic analysis (Gallo et al., 1996) which allows the distinction of state from action utterances. So the user can search for information on or from various aspects of an object.

References