Automatic indexing of multimedia documents
based on nominal phrases extraction

Abstract: Our research deals with the new technology context of communication and the numeric documents where text, image, and audio-video intermingle. For this point of view, a common character dealing with the research and to the access of information seem to be imposed. Thus, our objective of research is to reply to such request by studying on a variety of numeric documents in a goal of finding the strategies of content representation. This representation of the content can be only textual (full texts, abstracts, bibliographic notes incorporating abstracts, titles, etc.), then to apply a processing to index these multiform documents.

Our corpus which gathers multiform documents includes bibliographic notes of the National Institute of the Audiovisual (I.N.A. France). These notes have been elaborated by specialists of the documentation who had integrated text abstracts in these notes. Our contribution towards the area research on automatic indexing and particularly the indexing of multimedia documents which based on abstract contents.

Our study, firstly consists of studying the indexing documentary approaches. A state of the art is presented. Secondly, we present a particular approach based on Natural Language Processing (N.L.P.) compatible with our corpus. Finally, we will apply this approach in order to index these multiform documents by the use of their abstracts.

The used approach for the N.L.P. is the result of theoretical research by the SYDO Work-group in Lyon (SYDO: DOcumentary SYstem) based on Noun Phrases model (NP: French nominal phrases).

According to the SYDO approach, the implementation of the morpho-syntactic parser consists of the rewriting of the NP-grammar in a programming language and its application to the corpus to extract noun phrases.

The morphological categories are constructed so as to help in the elaboration of the rewriting NP-rules. The lexicon provides useful information for the morphological and the syntactical orders.

The objectives of the NP-model are:
- to allow the identification of NP,
- to determine the structure of the NP by creating relations between constituents and the storage of NP-representation in order to facilitate the information retrieval.
- to implement the mechanism of transition from the intensionnal logic (words that belong to the lexicon of the language) to the extensionnal logic (the noun phrase).

The grammar of recognition of the NP-structures is articulated around three levels:
- the Intensionnal level: PROPERTIES of the language (lexicon).
- the Intermediate level: the Transition from the intensionnal logic to the extensionnal logic.
- the Extensionnal level: OBJECTS of the real world (NP-structures).

Advances in Knowledge Organization, Vol.6(1998)