A Different Approach to UDC in a Mechanized Retrieval System

Attila Piros

University of Debrecen, Faculty of Informatics
atilla.piros@gmail.com

Abstract

Universal Decimal Classification (UDC) is one of the two big bibliographic and library classifications. It provides a coherent code system to collect and organize all branches of human knowledge in a hierarchical structure of concepts and a complex set of rules to create new concepts from the existing ones. As a formal language, it is capable to describe contents in a language-independent way.

Though it has been used worldwide for about a hundred years (different editions of it are available in more than 50 languages), there haven't been developed automatic or "mechanized" retrieval systems so far which could reflect and exploit the whole complexity of UDC indexes – partially because of the ambiguities of the language itself.

Most integrated library software systems handle UDC indexes just as simple keywords, what makes precise and efficient information retrieval impossible. The most sophisticated solutions can recognize some parts of UDC codes and copy these out of their context to a so-called KWOC index. Unfortunately, this method has several disadvantages; e.g. it leads to loose information about the context, result in ambiguous relationship between the concepts etc. – consequently, it raises the level of noise during the retrieval of data. The fact that not every part is added to the index and some or all special rules of UDC are avoided can also result in loosing information. Another problem is that the searcher must be thoroughly familiar with both the classification system and the subtleties of the software.

Using KWOC has been the only common and accepted solution to build indexes based on UDC for more then forty years, although modern descriptive and programming languages obviously make it possible to analyze and describe the inner structure of every simple and complex UDC index for using the resulting information during the search process. As a final result, the system would be able to retrieve, without significant noise, a complete set of records collected by exploiting the full capability of the relations and rules of UDC at high level, i.e. without any detailed knowledge of UDC by the

searcher. It means that the searcher doesn't even meet artificial UDC indexes – only the corresponding concepts expressed in natural language.

The first step in this way is to create a software interpreter to process UDC indexes. The author's goal is to present and demonstrate a working solution to this.

The presented interpreter gets a UDC index, its descriptions in different languages and the year of the UDC-version currently used. It parses the index and decomposes its notation which results in a hierarchical structure determined by the precedence of operators and auxiliaries.

One of the main advantages of the parser is that, in addition to the exact identification of the different parts of the index, it recognizes auxiliary types and their roles.

During the whole process, the interpreter takes the differences between UDC-versions into account as well – even when certain parts of the index analyzed belong to different versions, it is capable to produce adequate output.

As a consequence, the interpreter can validate UDC-indexes as well. If the index analyzed is invalid in the given UDC-version, the user will get an error message about the problem occurred.

Keywords: Universal Decimal Classification, Information Retrieval, Library Classification, Automatic Information Retrieval Systems, Integrated Library Software Systems