Automatic extraction of MeSH/DeCS terms in a health sciences repositories: the case of RUNA

Authors: Carmen Rodríguez Otero [1] , Emilio Lorenzo Gil [2], Adán Román Ruiz [2]

1. Bibliosaúde. Biblioteca Virtual do Sistema Sanitario Público de Galicia
2 Arvo Consultores y Tecnología

Introduction

To better manage and search the biomedical literature, the US National Library of Medicine developed the Medical Subject Heading (MeSH) controlled vocabulary. Thematic classification based on controlled vocabularies has been identified as a principal factor in document search and retrieval strategies. Unfortunately, given its specialized and manual nature, assigning MeSH terms to biomedical articles is a very complex, subjective, and time-consuming task, so Automated Keyword Extraction (AKE) systems are emerging as obvious solutions for its incorporation into systems that need to describe and manage thousands of documents, such as repositories.

The study shows the solution incorporated into RUNA, the institutional repository of the Public Health System of Galicia, in order to facilitate the subject classification on thematic vocabularies (MeSH-DeCS).

It specifically describes the system for the automatic extraction of document terms and how the solution has been integrated into the archive workflow in the repository in order to enable it to be complemented by expert cataloguers and therefore improve the quality of the thematic description carried out.

Methodology

The system constructed is integrated into the archive workflow of the repository documents with the aim of combining the advantages of automatic processing with the presence of an expert who undertakes the selection of the terms actually used. In this regard, the automatic extractor subsystem is seen as a pre-treatment of the document that proposes classification terms, which should then be validated and/or rejected by the repository's expert user.

Firstly, the document, which is normally in PDF or in other formats like WORD, is converted to a simple text format (txt). This step in the process does not only serve to normalise the document entry into the extractor system but also the transformed file is used by the full text indexer at the RUNA repository.

From this „simple" file, the initial selection of candidate terms is made, with the extraction of all the phrases, words, terms and concepts eligible to be descriptors.

This is followed by a process in which terms are assessed and selected. All the candidate terms are assessed by combining the properties of the terms (e.g. id they belong to the title) with machine learning techniques to ascertain the probability of the element being a key term. The system is configured to propose on the completion of this process a certain number of terms (10 in the RUNA repository). In the specific implementation carried out by the extraction engine, the elements extracted should belong to the MeSH-DeCS vocabulary.

The elements extracted are submitted to the cataloguer who based on their experience can accept or reject them or add new terms, as in a normal workflow process to the repository, thereby completing the document acceptance process in RUNA.

As a complementary aspect, the system initialises through the supply of a sufficient number of documents, in the form of a corpora, and its corresponding thematic metadating carried out by an human expert. The engine performs a first adjustment of the term probabilities, thereby implementing its initial learning.

Similarly, although this still has to be implemented in RUNA, the continuous flow of selections, reviews and approvals undertaken by the cataloguer may be used to feedback the extraction engine, evolving the probabilities assigned to each term and thereby improving the quality of the automatic proposals.

The solution described, together with the Dspace software of the RUNA repository, is based on *Maui*, a free software extractor (GPL licence). *Maui* is the acronym for Multi-purpose automatic topic indexing. The core of *Maui* is a machine-learning system called *WEKA*, which in turn incorporates the *AKE* algorithm for keyword extraction.

Results

The system constructed automises the extraction, description and indexing of the topical terms on the documents incorporated into the RUNA repository. As well as carrying out an initial automatic extraction, it enables a cataloguing expert to select (and add, if they decide so) the most appropriate MeSH-DeCS terms, thereby improving the quality and accuracy of the final cataloguing.

Conclusions

Automatic keyword extraction systems can be considered as a key complement that facilitates efficiently the accuracy of the thematic cataloguing of the documents incorporated into thematic repositories.