Ontology-driven Information System for the Knowledge Transmission in Prevention Education

Susanna Rivetti, Roberta Gazzarata, Mauro Giacomini
susanna.rivetti@libero.it
Department of Communication Computer and System Sciences, University of Genoa, Italy

Graz, July 24-27 2012

Abstract
Due to new medical discoveries in the last centuary, the culture of prevention has gained increasing importance in developed countries, but frequently, high quality information is submerged by many other unsafe sources. First of all, intelligent methods for the retrieval and synthesis of information are crucial for obtaining quick access to correct information during the education process. The retrieval component should make use of capabilities for generalization throughout a number of phases: indexing, retrieval, organization and presentation. Thus all of these tasks require an underlying concept-oriented approach which usually relies on ontological resources. A knowledge extraction tool has been developed for exploiting bilingual sources of domain knowledge, e.g. a subject reference system as a controlled language for document indexing and classification.

Material and Methods
Ontological resources
In the specific perspective of supporting linguistic inference in Information Extraction and Retrieval (IEW) for prevention education, the use of domain taxonomies as ontological resources is not straightforward. To accomplish the annotation task and the layout of a semantic search, we decided to develop a knowledge system based on the MeSH (Medical Subject Headings) tree.

Knowledge organization. Two databases have been built up: one to manage documents/links information, the other to manage the MeSH vocabulary. The first one contains all documents and links to external documents allowing managing the position of both of these in a hierarchic structure organized in four levels. In the second database we organized tables for concepts, terms, descriptor and qualifier with their specific information to implement an efficient way the tree and the semantic structures produced by NLM. The database contains English and Italian words. This collection of official Italian translations has been provided by the Italian Superior Sanitary Institute (ISS). The two databases are not directly linked to each other at a physical level, but they are linked in a logical way thanks to the developed application.

System description
The process consists of 3 main steps:
- documents/links included in the corpus were each annotated by a team of medical, health and safety experts that select and classify the information mainly contained within a certified organization web site; on the other hand, the IT experts that are specialist in data management, standard vocabulary and ontology. In this phase subjects are given to the document/link according to MeSH tree. The annotator facilitates the development of complicated annotation schemas and the annotated information can easily be modified in successive revisions. In this operation the authorized user has to choose one or more descriptors. Since MeSH defines a set of qualifiers available for each descriptor, if necessary, the user can also select one or more qualifiers for the considered descriptor from the corresponding set.

The chosen descriptors’ unique identifiers and the corresponding optional chosen qualifiers’ unique identifiers are then recorded in an ad hoc table within the first database, with the indication of the authorized user that suggested these subject.

Overview of the analysis pipeline. The heart of the application is an online search tool able to interpret user’s queries even in case these last have not been formulated using scientific medical terminology. So, we developed an algorithm that takes a sentence, in English or in Italian language, as input and produces a list of MeSH Metathesaurus concepts mapped in the sentence as output. The text analysis is handled through four different steps, scheduled in a pipeline-like architecture. The processing pipeline of our system is presented in Figure. Each user query is processed by the following modules:

1. The analysis starts with the detection of tokens. The module takes as input the query entered by the user, constituted by n words, and transforms it into an array of n-a words, where a indicates the number of words that represent stopwords (e.g. adverbs of time and place, interjections, articles, prepositions, numbers, conjunction, punctuation).

2. Experiments. The module returns directly the corresponding DescriptorUI. This stage has been made to make the search faster in case the user employs an appropriate technical language. Otherwise it excludes the words that are neither descriptors nor qualifiers, returning two arrays containing respectively: n-a-b-c couples DescriptorUI/DescriptorName and b qualifiers.

3. Each descriptor found in the previous passage are sought other descriptors containing it as part. The findings are saved in n-a-b-c tables.

4. This module checks if there are descriptors shared by more than a query word and sorts the found descriptors according to the number of query words involved.

Results retrieval and presentation. Once the algorithm has obtained the lists of DescriptorUI and QualifierUI as explained, the application queries the database for documents. If there are no documents in the database covering all the topics identified by the descriptors, the application selects those documents whose arguments more closely resemble the required ones, ordering them by number of matching descriptors. Once the topics have been found, in case the QualifierUI array is empty, the application returns the user all the documents that the database indicates. If, instead, the array of qualifier is not empty, the application queries the database again to find, among the documents obtained from the preceding interrogation, those who have the qualified indicators by the user.

References

PREDICTION AND RESEARCH