MAPi 2009-2013
Thesis proposal

Title

Semantic Indexing Approaches for Improved Information Retrieval – applications in scientific literature document databases

Background

The rapid increase in the amount of information available on the internet has created new challenges regarding the best ways to find the most relevant results for a user query. The current keyword-based search paradigm is rapidly reaching its limit and new forms of dealing with the large amount of (mostly) unstructured data available are needed. These rely, in one way or another, on natural language processing (NLP) methods to pre-process the data in order to achieve better results.

Similar problems as the ones encountered in general internet data are also encountered in the specific field of scientific literature databases. In this special case, the failure of current methods to accurately deal with the information growth means that a researcher looking for a specific subject may not find the most relevant results. That is, although more and more scientific knowledge is being produced and shared through publications, this increase is making it more difficult to be found by other researchers.

Aims

The general aims of the proposed work are to develop document indexing and retrieval methods that can be applied in the scientific literature field. These methods will be based on semantic indexing of scientific articles. This involves the analysis of documents in order to identify, annotate (using, for example, metadata and/or document markup) and index the occurrence of domain concepts like entities, terms, and relations between them. This analysis will be achieved through the application of text mining (TM) and NLP techniques and the use of available lexicons that define the domain entities and terms.

The expected results of this work consist of a set of tools and methods for semantic analysis and indexing of a set of documents. The tools should be flexible in order to receive a set of lexicons and rules, adapted to a specific domain, and produce a standardized annotation and indexation of the document set.

Supervisor

Sérgio Matos (aleixomatos@ua.pt)

Research Unit

UA/IEETA

A Research Grant may support this work in the context of the FCT project PTDC/EIA-CCO/100541/2008