Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology on the Web

Linda Fatima Soualmia\textsuperscript{1,2}, Catherine Barry\textsuperscript{3}, and Stefan J. Darmoni\textsuperscript{1,2}

\textsuperscript{1}CISMeF & L@STICS, Rouen University Hospital, 76031 Rouen, France
\{lina.soualmia, stefan.darmoni\}@chu-rouen.fr
\textsuperscript{2}PSI FRE CNRS 2645, INSA-Rouen, 76131 Mont-Saint Aignan, France
\textsuperscript{3}LaRIA, Picardie University, 80000 Amiens, France
barry@laria.u-picardie.fr

Abstract. This paper deals with the problem of information retrieval on the Web and present the CISMeF project (acronym of Catalogue and Index of French-speaking Medical Sites). Information retrieval in the CISMeF catalogue is done with a terminology that is similar to ontology of medical domain and a set of metadata. This allows us to place the project at an overlap between the present Web, which is informal, and the forthcoming Semantic Web. We also describe an ongoing work, which consists of applying three knowledge-based methods in order to enhance information retrieval.

1 Introduction

Nowadays the problematic is \textit{intelligent information retrieval} on the Web. The Semantic Web \cite{1} is an infrastructure that has to be built. It aims at creating a web where information semantics are represented in a form that can be understood by human as well as machines in order to enable computers and people to work in cooperation. One of its advantages is to bring sufficient information on the resources, by adding annotations in the form of \textit{metadata} and to describe formally and significantly their content according to an \textit{ontology}. Ontologies are considered to be powerful tools to lift ambiguity by providing a controlled vocabulary of terms and some specification of their meaning and are very useful for interoperability and for browsing and searching. Metadata describe Web information resources enhancing information retrieval.

In this paper we present the CISMeF\textsuperscript{1} project \cite{2} (acronym of Catalogue and Index of French-speaking Medical Sites) developed since 1995. The objective of CISMeF is to help health professionals, as well as students and the general public, during their search for electronic health information. The CISMeF catalogue describes and indexes a large number of health information resources \((n=11,504)\). A resource can be a Web site, Web pages, documents, reports and teaching material: any support that may contain health information. The resources are selected according to strict criteria by the librarian team and are indexed according to a methodology. The resources indexed in the CISMeF catalogue are described according to a terminology that is similar to an ontology of the medical domain, and a set of metadata elements. This structure en-

\footnote{1 http://www.chu-rouen.fr/cismef/}

© Springer-Verlag Berlin Heidelberg 2003
ables us to place the project at an overlap between the present informal Web, mainly composed by HTML pages, and the forthcoming Semantic Web. We also describe in this paper an ongoing work which consists of applying three knowledge-based methods (natural language processing, knowledge discovery in databases and reasoning on ontologies) to enhance information retrieval into CISMeF.

2 Towards a Medical Semantic Web

Metadata is data about data and specifically in the context of the Web, it is data that describe Web resources. When properly implemented, metadata can enhance information retrieval. In CISMeF several sets of metadata elements are used. The resource indexed are described the Dublin Core (DC) elements set [3] (e.g. author, date). DC is not a complete solution, it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF [2](e.g. institution, target public). Two additional fields are in the resources intended for the health professionals: indication of the evidence-based medicine and the method used to determine it. In the teaching resources eleven elements of the IEEE 1484 LOM (Learning Object Metadata) “Educational” category are added. The metadata format was the HTML language in 1995. Since December 2002, the format used is RDF, a Semantic Web language, within the ongoing MedCIRCLE project [4], developed to qualify health information quality.

The catalogue resources are indexed according to the CISMeF terminology, which is based on the MeSH [5] concepts and its French translation. We have not used the UMLS [6] because there is no available French translation. Approximately 22,000 keywords (e.g. hepatitis) and 84 qualifiers (e.g. complications) compose the MeSH thesaurus, in its 2003 version. These concepts are organized into hierarchies going from the most general on the top to the most specific in the bottom of the hierarchy. The qualifiers, organized into hierarchies, specify which particular aspect of a keyword is addressed. The keywords and the qualifiers that are dispersed in several trees but are semantically related in CISMeF are gathered according to metaterms (n=66). They concern medical specialties. In addition, a hierarchy of resource types (n=127) describes the nature of the resource (e.g. clinical guidelines). The metaterms and resource types enhance information retrieval into the catalogue when searching for “guidelines in cardiology”, where cardiology is a metaterm and guidelines is a resource type, which is not possible using the MeSH thesaurus.

The CISMeF terminology has the same structure as a terminological ontology [7]. The vocabulary describes major terms of the medical domain and is well known by the librarians and the health professional. Each concept has a preferred term to express it in natural language, a set of properties, a natural language definition that allows to differentiate it from the concepts it subsumes and those that subsume it, a set of synonyms and a set of rules and constraints.

3 Enhancing Information Retrieval

The submitted queries over the search engine are seldom matched to the terms of the ontology. We have extracted and analyzed 1,552,776 queries of the http server log