Chapter VII
TtoO:
Mining a Thesaurus and Texts to Build and Update a Domain Ontology

Josiane Mothe
Institut de Recherche en Informatique de Toulouse Université de Toulouse, France

Nathalie Hernandez
Institut de Recherche en Informatique de Toulouse Université de Toulouse, France

ABSTRACT

This chapter introduces a method re-using a thesaurus built for a given domain, in order to create new resources of a higher semantic level in the form of an ontology. Considering ontologies for data-mining tasks relies on the intuition that the meaning of textual information depends on the conceptual relations between the objects to which they refer rather than on the linguistic and statistical relations of their content. To put forward such advanced mechanisms, the first step is to build the ontologies. The originality of the method is that it is based both on the knowledge extracted from a thesaurus and on the knowledge semiautomatically extracted from a textual corpus. The whole process is semiautomated and experts' tasks are limited to validating certain steps. In parallel, we have developed mechanisms based on the obtained ontology to accomplish a science monitoring task. An example will be given.

INTRODUCTION

Scientific and technical data, whatever their types (textual, factual, formal, or nonformal), constitute strategic mines of information for decision makers in economic intelligence and science monitoring activities, and for researchers and engineers (e.g., for scientific and technological watch). However, in front of the growing mass of information, these activities require increasingly powerful systems offering greater possibilities for exploring and representing the collected information or extracted

knowledge. Upstream, they must ensure search, selection and filtering of the electronic information available. Downstream, when communicating and restituting results, they must privilege ergonomics in presentation, exploration, navigation, and synthesis.

To be manipulated by advanced processes, textual document content has first to be represented synthetically. This process is known as indexing and consists in defining a set of terms or descriptors that best correspond to the text content. Descriptors can either be extracted from the document themselves or by considering external resources such as thesauri. In the first approach, which can be fully automatic and thus more appropriate for large volumes of electronic documents, the texts are analyzed and the most representative terms are extracted (Salton, 1971). These are the terms which constitute the indexing language. The automatic weighting of index terms (Robertson & Sparck-Jones, 1976), their stemming (Porter, 1981), the automatic query reformulation by relevance feedback (Harman, 1992) or per addition of co-occurring terms (Qiu & Frei, 1993) in information retrieval systems (IRS) are methods associated with automatic indexing and have been effective as attested by international evaluation campaigns, such as text retrieval conference (trec.nist.gov). One common point of all these approaches is that they make the assumption that the documents contain all the knowledge necessary for their indexing. On the other hand, thesauri are used to control the terminology representing the documents by translating the natural language of the documents into a stricter language (documentary language) (Chaumier, 1988). A thesaurus is based on a hierarchical structuring of one or more domains of knowledge in which terms of one or more natural languages are organized by relations between them, represented with conventional signs (AFNOR, 1987). With ISO 2788 and ANSI Z39, their contents can be standardized in terms of equivalence, hierarchical, and associative relations between lexemes. Indexing based on a thesaurus is generally carried out manually by librarians who, starting from their expertise, choose the terms of the thesaurus constituting the index of each document read. In an IRS, the same thesaurus is then used to restrict the range of a query or, on the contrary, to extend it, according to the needs of the user and the contents of the collection. Other types of systems combine the use of a thesaurus with classification and navigation mechanisms. The Cat a Cone system (Hearst, 1997) or IRAIA system (Englmeier & Mothe, 2003) make use of the hierarchical structure of the thesaurus to allow the user to browse within this structure and thus to access the documents associated with the terms. Compared to automatic indexing, the thesaurus approach leads to more semantic indexing, as terms are considered within their context (meaning and related terms). However, using thesauri raises several problems: they are created manually and their construction requires considerable effort; their updating is necessary; their format is not standardized (ASCII files, HTML, data bases coexist); finally, thesauri have a weak degree of formalization since they are built to be used by domain experts and not by automatic processes. Various solutions to these issues have been proposed in the literature. Automatic thesauri construction can call upon techniques based on term correlation (Tudhope, 2001), document classification (Crouch & Yang, 1992), or natural language processing (Grefenstette, 1992). On the other hand, the standards under development within the framework of the W3C as SKOS Core (Miles et al., 2005) aim at making the thesaurus migrate towards resources that are more homogeneous and represented in a formal way by using OWL language (McGuiness & Harmelen, 2004) and making these resources available on the Semantic Web. However, thesauri represent a domain in terms of indexing categories and not in terms or meaning. They do not have a level of conceptual abstraction (Soergel et al., 2004), which plays a crucial role in man-machine communication. Ontologies make it possible to