IRM PRESS
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB9196

Chapter III

Interactive Indexing of Documents with a Multilingual Thesaurus

Ulrich Schiel
Universidade Federal de Campina Grande, Brazil

Ianna M.S.F. de Sousa
Universidade Federal de Campina Grande, Brazil

ABSTRACT

With the growing significance of digital libraries and the Internet, more and more electronic texts become accessible to a wide and geographically disperse public. This requires adequate tools to facilitate indexing, storage and retrieval of documents written in different languages. We present a method for semi-automatic indexing of electronic documents and construction of a multilingual thesaurus, which can be used for query formulation and information retrieval. We use special dictionaries and user interaction in order to solve ambiguities and find adequate canonical terms in the language and an adequate abstract language-independent term. The abstract thesaurus is updated incrementally by new indexed documents and is used to search for documents using adequate terms.

INTRODUCTION

The growing relevance of digital libraries is generally recognized (Haddouti, 1997). A digital library typically contains hundreds or thousands of documents. It is also

recognized that, even though English is the dominant language, documents in other languages are of great significance and, moreover, users want to retrieve documents in several languages associated to a topic, stated in their own language (Haddouti, 1997; Go02). This is especially true in regions such as the European Community or Asia. Therefore a multilingual environment is needed to attend user requests to digital libraries.

The multilingual communication between users and the library can be realized in two ways:

- The user query is translated to the several languages of existing documents and submitted to the library.
- The documents are indexed and the extracted terms are converted to a language-neutral thesaurus (called multilingual thesaurus). The same occurs with the query, and the correspondence between query terms and documents is obtained via the neutral thesaurus.

The first solution is the most widely used in the Cross-Language Information Retrieval (CLIR) community (Go02; Ogden & Davis, 2000; Oard, 1999). It applies also to other information retrieval environments, such as the World Wide Web. For digital libraries, with thousands of documents, indexing of incoming documents and a good association structure between index terms and documents can become crucial for efficient document retrieval.

In order to get an extensive and precise retrieval of textual information, a correct and consistent analysis of incoming documents is necessary. The most broadly used technique for this analysis is indexing. An index file becomes an intermediate representation between a query and the document base.

One of the most popular structures for complex indexes is a semantic net of lexical terms of a language, called thesaurus. The nodes are single or composed terms, and the links are pre-defined semantic relationships between these terms, such as synonyms, hyponyms and metonyms.

Despite that the importance of multilingual thesauri has been recognized (Go02), nearly all research effort in Cross-Lingual Information Retrieval has been done on the query side and not on the indexing of incoming documents (Ogden & Davis, 2000; Oard, 1999; Haddouti, 1997).

Indexing in a multilingual environment can be divided in three steps:

1. language-dependent canonical term extraction (including stop-word elimination, stemming, word-sense disambiguation);
2. language-neutral term finding; and
3. update of the term-document association lattice.

Bruandet (1989) has developed an automatic indexing technique for electronic documents, which was extended by Gammoudi (1993) to optimal thesaurus generation for a given set of documents. The nodes of the thesaurus are bipartite rectangles where the left side contains a set of terms and the right side the set of documents indexed by the terms. Higher rectangles in the thesaurus contain broader term sets and fewer documents. One extension to this technique is the algorithm of Pinto (1997), which permits an incremental addition of index terms of new incoming documents, updating the thesaurus.