Minimum Redundancy Cut in Ontologies for Semantic Indexing*

Florian Seydoux and Jean-Cédric Chappelier

School of Computer and Communication Sciences,
École Polytechnique Fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
{florian.seydoux, jean-cedric.chappelier}@epfl.ch

Abstract. This paper presents a new method that aims at improving semantic indexing while reducing the number of indexing terms. Indexing terms are determined using a minimum redundancy cut in a hierarchy of conceptual hypernyms provided by an ontology (e.g. *WordNet*, *EDR*). The results of some information retrieval experiments carried out on several standard document collections using the *EDR* ontology are presented, illustrating the benefit of the method.

1 Introduction

The main idea of semantic indexing is to use word senses rather than, or in addition to the words[1] for indexing documents, in order to improve both recall (by handling synonymy) and precision (by handling homonymy and polysemy). However, the related experiments reported in the Information Retrieval (IR) literature lead to contradicting results: some claim that this substitution (or addition), carried out in an automatic way, degrades the performances [1–4]; for others conversely, the gain seems significant [5–9].

Although it seems desirable for document indexing to take a maximum of semantic information into account, the resulting expansion of the data processed could happen to be counter-productive. Indeed, the growth of the number of indexing terms not only increases the processing time, but could also reduce the precision: discriminating documents by using a very large number of indexing terms is a hard task ("*curse of dimensionality*" effect). This problem is not new, and various techniques aiming at reducing the size of the indexing set already exist: filtering by stoplist, part of speech tags, frequencies, or through statistical techniques such as LSI [10] or PLSI [11]. However, most of these techniques are not adapted to the case where an explicit semantic information is available in the form of an ontology, i.e. with some underlying formal – not statistical – structure.

The focus of the work presented here is to use ontologies to create semantic indexing sets of "sensible" sizes. This relates, but from a different point of

* This work was partially supported by the Swiss National Fund for Scientific Research (SNFSR) under grant #200020–103529.

[1] Usually stems or lemmas.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 658–668, 2005.

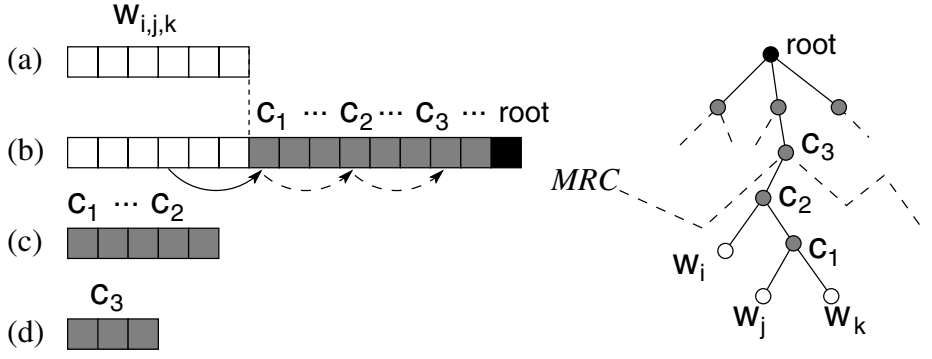

Fig. 1. Different indexing scheme: (a) usual indexing with words (stems or lemmas); (b) using a semantic ontology (displayed on the right), each indexing term is extended with all the concepts that dominate it; this leads to an explosion of the number of indexes for documents; (c) synset (or hypernyms synsets) indexing: each indexing term is replaced with its (hypernyms) synset; this, in principle, can reduce the size of the indexing set since all the indexing terms that are covered by the same hypernym are regrouped in one single indexing feature; (d) Minimum Redundancy Cut (*MRC*) indexing: each indexing term is replaced with its dominating concept defined by *MRC*. This reduces further the size of the indexing set since all the indexing terms that are subsumed by the same concept in the *MRC* are regrouped in one single indexing feature.

view, with experiments reported in [8], [12] or [9], which uses the synsets (or hypernyms synsets [9]) of *WordNet* as indexing terms. Our work goes one step ahead, selecting the indexing set using an information theory based criterion, the *Minimum Redundancy Cut* (*MRC*, see Fig. 1). This criterion is applied to the inclusive "is-a" relation (hypernyms) provided by the *EDR* taxonomy [13].

2 Ontology-Cut Model

2.1 Goals

The choice of the appropriate hypernym (a concept in the ontology) to be used for representing a word is not easy: be it too general, the performances of the system will degrade (lack of precision); be it too specific, the indexing set will not reduce enough, preserving some distinction between words with close senses (lack of recall).

To select the appropriate level of conceptual indexing, we consider cuts in the ontology. A cut is a minimal set[2] of nodes in the ontology defining a coverage of all the leaf nodes (i.e. words). Each node in the cut is used to represent every leaf node it dominates.

The problem is to find a computable strategy to select an optimal cut. For a related task, Li and Abe [14] use the Minimum Description Length principle

[2] By "minimal set" we mean that no node can be removed from the set without decreasing it's coverage (i.e. the number of leaves it dominates).