

Approche évolutive des notions de base pour une représentation thématique des connaissances générales

Alain Joubert, Mathieu Lafourcade, Didier Schwab

LIRMM – UM2

Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier

161 rue Ada, 34392 Montpellier Cédex 5, FRANCE

{[alain.joubert](mailto:alain.joubert@lirmm.fr), [lafourcade](mailto:lafourcade@lirmm.fr), [schwab](mailto:schwab@lirmm.fr)}@lirmm.fr

Résumé

Dans le domaine du Traitement Automatique du Langage Naturel, pour élaborer un système de représentation thématique des connaissances générales, des méthodes s'appuyant sur des thésaurus sont utilisées depuis une quinzaine d'années. Un thésaurus est constitué d'un ensemble de concepts qui définissent un système générateur d'un espace vectoriel modélisant les connaissances générales. Ces concepts, souvent organisés en une hiérarchie arborescente, constituent un instrument fondamental, mais totalement figé. Même si les notions évoluent (nous pensons par exemple aux domaines techniques), un thésaurus ne peut quant à lui être modifié que lors d'un processus particulièrement lourd, car nécessitant la collaboration d'experts humains. C'est à ce problème que nous nous attaquons ici. Après avoir détaillé les caractéristiques que doit posséder un système générateur de l'espace vectoriel de modélisation des connaissances, nous définissons les « notions de base ». Celles-ci, dont la construction s'appuie initialement sur les concepts d'un thésaurus, constituent un autre système générateur de cet espace vectoriel. Nous abordons la détermination des acceptions exprimant les notions de base, ce qui nous amène naturellement à nous poser la question de leur nombre. Enfin, nous explicitons comment, s'affranchissant des concepts du thésaurus, ces notions de base évoluent par un processus itératif au fur et à mesure de l'analyse de nouveaux textes.

Mots clés : thésaurus, vecteurs conceptuels, notions de base, évolutivité

Abstract

In the field of Natural Language Processing, in order to work out a thematic representation system of general knowledge, methods leaning on thesaurus are used since about fifteen years. A thesaurus consists of a set of concepts which define a generating system of a vector space modelling general knowledge. These concepts, often organized in an arborescent hierarchy, constitute a fundamental, but completely fixed tool. Even if the concepts evolve (we think for example of the technical fields), a thesaurus can evolve as for him only at the time of a particularly heavy process, because requiring the collaboration of human experts. After we detailed the characteristics which a generating system of the vector space of knowledge modelling must have, we define the "basic notions". Those, whose construction is based initially on the concepts of a thesaurus, constitute another generating system of this vector space. We approach the determination of the acceptions expressing the basic notions, which naturally leads us to ask the question of their number. Lastly, we clarify how, being freed from the concepts of the thesaurus, the basic notions evolve by an iterative process progressively with the analysis of new texts.

Keywords : thesaurus, conceptual vectors, basic notions, evolution