Mathématiques et sciences humaines
Mathematics and social sciences

148 | Hiver 1999
Varia

Apprentissage d'un ensemble pré-structuré de concepts d'un domaine : l'outil Galex

Learning of the pre-structured concept set of a domain: the Galex tool

Nicolas Turenne

Édition électronique
URL : http://journals.openedition.org/msh/2788
DOI : 10.4000/msh.2788
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 1 décembre 1999
ISSN : 0987-6936

Référence électronique
Nicolas Turenne, « Apprentissage d'un ensemble pré-structuré de concepts d'un domaine : l'outil Galex », *Mathématiques et sciences humaines* [En ligne], 148 | Hiver 1999, mis en ligne le 10 février 2006, consulté le 06 mai 2019. URL : http://journals.openedition.org/msh/2788 ; DOI : 10.4000/msh.2788

Apprentissage d'un ensemble pré-structuré de concepts d'un domaine : l'outil Galex

Learning of the pre-structured concept set of a domain: the Galex tool

Nicolas Turenne

RÉSUMÉS

La quantité d'information textuelle augmente de façon exponentielle aussi bien comme archives que documents de travail dans les organisations académiques, dans les administrations et dans les entreprises. Une solution pour structurer cette montagne de données textuelles est de construire un modèle de connaissances pour indexer cette information. L'acquisition de connaissances doit permettre d'extraire et classifier les données pour aboutir à une indexation conceptuelle. Traditionnellement les méthodes de classification d'analyse de données étaient adaptées pour des tables classiques de données de la forme objet/attribut/valeur. Nous présentons Galex (Graph Analyzer for LEXicometry) qui développe une structuration de la connaissance grâce à une méthode de clustering de termes. Cette structuration a pour but de synthétiser le contenu d'information présentant un intérêt majeur dans des applications de filtrage d'information ou de navigation hypertextuelle sur des documents similaires. Galex prend en compte la nature des données sur lesquelles il s'applique : le langage naturel. La complexité du langage naturel est bien connue : ambiguïté de sens, constructions grammaticales multiples de la phrase, style, création de termesá Nous montrons qu'à travers l'intégration de notions mal définies mais utiles telles que "concept", "ontologie" et "corpus", le clustering peut être amélioré par adjonctions de connaissances linguistiques. Nous basons notre approche sur des phénomènes

typiques tels que des relations graphe-statistiques entre termes, des relations de schéma dans un contexte et la réduction canonique de formes variantes.

The huge amount of electronic textual information increases exponentially just as easily as archives and working documents in academic organizations, in administration and in firms. A solution for structuring this mountain of textual database is to build a knowledge model to index this information. One way can be obtained by data extraction and classification producing conceptual indexing by knowledge acquisition. Traditionally the classification methods of Data Analysis were adapted while used for the classical table of data under an object/characteristics/ value format. We present Galex (Graph Analyzer for LEXicometry) which develops structuration of knowledge by a term clustering method. This structuration synthetizes the content of information providing the mapping data to information filtering or hypertextual navigation on similar documents. Galex aims at taking into account the nature of the data to which it is applied : natural language. The complexity of natural language is well known: sense ambiguity, multiple grammatical construction of sentence, style, term creationáWe show through integration of poorly defined, though useful as concept, ontology, term and corpus, notions that clustering can be improved by adding linguistic knowledge. We base our approach on typical phenomena such as graph-statistical relations between terms, scheme relations in a context and canonical reduction of variants.

INDEX

Mots-clés : acquisition de connaissances, analyse de corpus, analyse de données, apprentissage de concepts, clustering de termes, fouille de texte, ontologie, text-mining
Subjects : cognitive sciences, computer sciences, data analysis, graphs, linguistics
Thèmes : cognitives (sciences), données (analyse des), graphes, informatique, linguistique
Keywords : concept learning, corpus analysis, knowledge acquisition, ontology, statistical data analysis, terms clustering, text-mining