Semantic Indexing Improvement for Textual Documents: Contribution of Classification by Fuzzy Association Rules

Authors : Mohsen Maraoui

Abstract : In the aim of natural language processing applications improvement, such as information retrieval, machine translation, lexical disambiguation, we focus on statistical approach to semantic indexing for multilingual text documents based on conceptual network formalism. We propose to use this formalism as an indexing language to represent the descriptive concepts and their weighting. These concepts represent the content of the document. Our contribution is based on two steps. In the first step, we propose the extraction of index terms using the multilingual lexical resource Euro WordNet (EWN). In the second step, we pass from the representation of index terms to the representation of index concepts through conceptual network formalism. This network is generated using the EWN resource and pass by a classification step based on association rules model (in attempt to discover the non-taxonomic relations or contextual relations between the concepts of a document). These relations are latent relations buried in the text and carried by the semantic context of the co-occurrence of concepts in the document. Our proposed indexing approach can be applied to text documents in various languages because it is based on a linguistic method adapted to the language through a multilingual thesaurus. Next, we apply the same statistical process regardless of the language in order to extract the significant concepts and their associated weights. We prove that the proposed indexing approach provides encouraging results.

Keywords : concept extraction, conceptual network formalism, fuzzy association rules, multilingual thesaurus, semantic indexing

1

Conference Title : ICALT 2018 : International Conference on Advanced Learning Technologies **Conference Location :** Paris, France

Conference Dates : October 29-30, 2018