Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach

I. Alonso, D. Contreras

Abstract— One promise of current information retrieval systems is the capability to identify risk groups for certain diseases and pathologies based on the automatic analysis of vast amounts of Electronic Medical Records repositories. However, the complexity and the degree of specialization of the language used by the experts in this context, make this task both challenging and complex. In this work, we introduce a novel experimental study to evaluate the performance of the two semantic similarity metrics (Path and Intrinsic IC-Path, both widely accepted in the literature) in a real-life information retrieval situation. In order to achieve this goal and due to the lack of methodologies for this context in the literature, we propose a straightforward information retrieval system for the biomedical field based on the UMLS Metathesaurus and on semantic similarity metrics. In contrast with previous studies which focus on testbeds with limited and controlled sets of concepts, we use a large amount of information (101,712 medical documents extracted from TREC Medical Records Track 2011). Our results show that in real-life cases, both metrics display similar performance, Path (F-Measure = 0.430) e Intrinsic IC-Path (F-Measure = 0.427). Thereby we suggest that the use of Intrinsic IC-Path is not justified in real scenarios.

Index Terms— Semantic similarity; Information retrieval; Electronic Health Record; UMLS

Due to copyright restriction we cannot distribute this content on the web. However, clicking on the next link, authors will be able to distribute to you the full version of the paper:
Request full paper to the authors

If you institution has a electronic subscription to Expert Systems with Applications, you can download the paper from the journal website:
Access to the Journal website