

The 13<sup>th</sup> International Scientific Conference  
eLearning and Software for Education  
Bucharest, April 27-28, 2017  
10.12753/2066-026X-17-177

**SEMANTIC ANNOTATION AND AUTOMATED TEXT CATEGORIZATION  
USING COHESION NETWORK ANALYSIS**

Gabriel GUTU, Mihai DASCALU

*Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania  
{gabriel.gutu, mihai.dascalu}@cs.pub.ro*

Dominic HEUTELBECK, Matthias HEMMJE

*FTK, Germany  
dheutelbeck}@ftk.de, mhemmje@ftk.de*

Wim WESTERA

*Open University of the Netherlands, the Netherlands  
wim.westera@ou.nl*

Stefan TRAUSAN-MATU

*Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania  
stefan.trausan@cs.pub.ro*

**Abstract:** *With the increasing amount of published scientific papers, it becomes paramount for learners and researchers alike to use tools that semantically annotate resources in order to facilitate the information retrieval process. Thus, we introduce a semantic annotation tool incorporated within our ReaderBench framework to provide recommendations regarding categories that should be used for automated labelling. Currently, the tool categorizes input documents based on the ACM Computing Classification System ([http://dl.acm.org/ccs\\_flat.cfm](http://dl.acm.org/ccs_flat.cfm)) taxonomy from 2012. The Semantic Annotation tool provides also suggestions and cohesion scores for the most relevant keywords covered by the paper, allowing researchers to automatically extract the paper's topics. Therefore, the semantic annotation algorithms involved within the tool ensure a cohesion-centered and in-depth representation of discourse. The underlying adaptive technologies support academia with potential suggestions of automated categorization and keywords generation useful when submitting scientific papers or properly assigning papers for review. A more specific objective is to facilitate the classification of publications in the internal Digital Library of the RAGE project meant to support researchers from the RAGE ecosystem. Additionally, the Semantic Annotation tool provides cohesion scores between the abstract, authors' keywords and the entire paper's textual content. These scores may provide useful insights in terms of generating personalized recommendations of keywords that are representative for an article, or recommendations for rewriting the abstract in a cohesive manner in accordance with the entire paper. To this aim, the SemEval corpora comprising of 244 scientific papers classified into four of the ACM CCS 1998 categories was used to validate our tool. Hence, we applied a clustering algorithm to group semantically related papers and compared the generated clusters with the initial group assignments.*

**Keywords:** *Semantic Annotation; Text Cohesion; Text Categorization; Topic Mining; Discourse Analysis; Adaptive Technologies; Inter-disciplinary Research.*

## I. INTRODUCTION

The exponential growth of the number of available digital documents increased the preoccupation for automated text categorization systems within the last years. These systems are