Annals of the Academy of Romanian Scientists
Series on Science and Technology of Information
ISSN 2066-8562 Volume **1**, Number **1**/**2008**

TEXT CLASSIFICATION IN ROMANIAN FOR AN INTELLIGENT NEWS PORTAL

Traian REBEDEA[1], Ştefan TRĂUŞAN-MATU[2]

Rezumat. *Articolul prezintă un modul de clasificare a textului pentru un portal de ştiri în limba română. Tehnicile de prelucrare statistică a limbajului natural sunt combinate cu scopul de a realiza funcţionarea complet autonomă a portalului. Fiecare ştire este colectată automat dintr-un mare număr de surse de ştiri folosind clasificarea web. De aceea, tehnicile de îmbogăţire a cunoştinţelor programului sunt folosite pentru clasificarea automată a fluxului ştirilor. Mai întâi, termenii sunt clasificaţi folosind un algoritm aglomerativ răsturnat; clasele rezultate corespund subiectelor principale. De aceea, sunt preluate mai multe informaţii despre fiecare dintre subiectele principale din mai multe surse de ştiri. Apoi, algoritmii de clasificare a textului sunt aplicaţi pentru etichetarea automată a fiecărei clase de ştiri (al căror număr este predeterminat). Au fost folosite peste o mie de ştiri, atât pentru antrenarea cât şi pentru evaluarea clasificatorilor. În articol este prezentată o comparare completă a rezultatelor obţinute prin fiecare metodă. Mai mult, sunt prezentate problemele specifice care apar datorită particularităţilor limbii române şi sunt discutate soluţiile găsite.*

Abstract. *The paper presents a text classification module for a news portal for the Romanian language. Statistical natural language processing techniques are combined in order to achieve a completely autonomous functionality of the portal. The news items are automatically collected from a large number of news sources using web syndication. Afterward, machine-learning techniques are used for achieving an automatic classification of the news stream. Firstly, the items are clustered using a bottom-up agglomerative algorithm and the resulting groups correspond to the main news topics. Thus, more information about each of the main topics is acquired from various news sources. Secondly, text classification algorithms are applied to automatically label each group of news items in a predetermined number of classes. More than a thousand news items were employed for both the training and the evaluation of the classifiers. The paper presents a complete comparison of the results obtained for each method. Moreover, specific problems that arose due to the particularities of the Romanian language are presented and the solutions found are discussed.*

Keywords: natural language processing, text clustering, classification, news portal, intelligent agent

[1]Drd. Eng., University "Politehnica" Bucharest, Department of Computer Science and Engineering.
[2]Prof. Dr. Eng., University "Politehnica" Bucharest, Department of Computer Science and Engineering; corresponding member of the Academy of Romanian Scientists (trausan@gmail.com).