Evaluation of Internal Validity Measures in Short-Text Corpora*

Diego Ingaramo[1], David Pinto[2,3], Paolo Rosso[2], Marcelo Errecalde[1]

[1]Development and Research Laboratory in Computacional Intelligence (LIDIC),
UNSL, Argentina
[2]Natural Language Engineering Lab.,
Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain
[3]Faculty of Computer Science (FCC),
BUAP, Mexico
{*daingara,merreca*}*@unsl.edu.ar,* {*prosso,dpinto*}*@dsic.upv.es*

Abstract. Short texts clustering is one of the most difficult tasks in natural language processing due to the low frequencies of the document terms. We are interested in analysing these kind of corpora in order to develop novel techniques that may be used to improve results obtained by classical clustering algorithms. In this paper we are presenting an evaluation of different internal clustering validity measures in order to determine the possible correlation between these measures and that of the F-Measure, a well-known external clustering measure used to calculate the performance of clustering algorithms. We have used several short-text corpora in the experiments carried out. The obtained correlation with a particular set of internal validity measures let us to conclude that some of them may be used to improve the performance of text clustering algorithms.

1 Introduction

Document clustering consists in the assignment of documents to unknown categories. This task is more difficult than supervised text categorization [13,8] because the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that clustering results cannot be evaluated with typical external measures like F-Measure and, therefore, the quality of the resulting groups is evaluated with respect to *structural properties* or *internal measures*. Classical internal measures used as cluster validity measures include the *Dunn* and *Davies-Bouldin* indexes, new graph-based measures like *Density Expected Measure* and Λ-Measure as well as some measures based on the corpus vocabulary overlapping.

When clustering techniques are applied to collections containing *very short* documents, additional difficulties are introduced due to the low frequencies of

* This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant