Clustering Abstracts of Scientific Texts using the Transition Point Technique *

David Pinto[1,2], Héctor Jiménez-Salazar[1], and Paolo Rosso[2]

[1] Faculty of Computer Science, BUAP, Puebla 72570,
Ciudad Universitaria, MEXICO
{davideduardopinto, hgimenezs}@gmail.com

[2] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, SPAIN
{dpinto, prosso}@dsic.upv.es

Abstract. Free access to scientific papers in major digital libraries and other web repositories is limited to only their abstracts. Current keyword-based techniques fail on narrow domain-oriented libraries, e.g., those containing only documents on high energy physics like those of the *hep-ex* collection of CERN. We propose a simple procedure to cluster abstracts which consists in applying the transition point technique during the term selection process. This technique uses the mid-frequency terms to index the documents due to the fact that they have a high semantic content. In the experiments we have carried out, the transition point approach has been compared with well known unsupervised term selection techniques. Transition point technique shown that it is possible to obtain a better performance than traditional methods. Moreover, we propose an approach to analyse the stability of transition point term selection method.

1 Introduction

Nowadays, very short text clustering on narrow domains has not received too much attention by the computational linguistic community. This is derived from the high challenge that this problem implies, since the obtained results are very unstable or imprecise when clustering abstracts of scientific papers, technical reports, patents, etc. But, as we can see, most digital libraries and other web-based repositories of scientific and technical information nowadays provide free access only to abstracts and not to the full texts of the documents. Moreover, some institutions, like the well known CERN[1], receive hundreds of publications every day that must be categorized on some specific domain with an unknown

* This work was partially supported by BUAP-VIEP 3/G/ING/05, R2D2 (CICYT TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), and Generalitat Valenciana Grant (CTESIN/2005/012).

[1] Centre Européen pour la Recherche Nucléaire