Sampling and Feature Selection in a Genetic Algorithm for Document Clustering

A. Casillas[1], M. T. González de Lena[2], and R. Martínez[2]

[1] Dpt. Electricidad y Electrónica
Universidad del País Vasco
arantza@we.lc.ehu.es

[2] Dpt. Informática, Estadística y Telemática
Universidad Rey Juan Carlos
{m.t.gonzalez,r.martinez}@escet.urjc.es

Abstract. In this paper we describe a Genetic Algorithm for document clustering that includes a sampling technique to reduce computation time. This algorithm calculates an approximation of the optimum k value, and solves the best grouping of the documents into these k clusters. We evaluate this algorithm with sets of documents that are the output of a query in a search engine. Two types of experiment are carried out to determine: (1) how the genetic algorithm works with a sample of documents, (2) which document features lead to the best clustering according to an external evaluation. On the one hand, our GA with sampling performs the clustering in a time that makes interaction with a search engine viable. On the other hand, our GA approach with the representation of the documents by means of entities leads to better results than representation by lemmas only.

1 Introduction

Clustering involves dividing a set of n objects into a specified number of clusters k, so that objects are similar to other objects in the same cluster, and different from objects in other clusters. Clustering algorithms can work with objects of different kinds, but we have focused on documents.

Several clustering approaches assume that the appropriate value of k is known. However, there may be numerous situations in which it is not possible to know the appropriate number of clusters, or even an approximation. For instance, if we want to divide into clusters a set of documents that are the result of a query to a search engine, the value of k can change for each set of documents that results from interaction with the engine. Amongst the first to recommend that automatic clustering might prove useful in document retrieval were [Good 58], [Fairthorne 61], and [Needham 61].

In our first approach [Casillas et al. 03], we dealt with the problem of clustering a set of documents without prior evidence on the appropriate number of clusters. Our main aim was to provide an approximation of an appropriate value of k, with an acceptable computation cost, for a small number of documents.