Chapter 10
E-Government Documents and Data Clustering

Goran Šimić
University of Defense, Serbia

ABSTRACT

This chapter is about documents and data clustering as a process of preparing the information resources stored in the e-government systems for advanced search. These resources are mainly represented as textual data stored as field values in the databases or located as documents in file repositories. Due to their growth in number, search for some specific information takes more time. Different techniques are used for this purpose. Most of them include information retrieval based on a variety of text similarity measures. The cost of such processing depends on preparation of resources for searching. Clustering represents the most commonly used technique for such a purpose, and this fact is the basic motive for this chapter.

INTRODUCTION

According to complexity of contemporary life, the governments try to establish close relations with citizens as much as possible in order to better understand their needs and problems, to offer them different kind of help and to satisfy their expectations. In such a situation the institutions offer the variety of Web services named e-government, or open government to make the information their systems already hold accessible to the people regardless of time and location. The information retrieval (IR) from big data collections stored in e-government information systems (IS) represents the important part of the solution. The data in such collections are heterogeneously structured and presented. Therefore, they can be hardly categorized depending on information they contain. The clustering represents the way for grouping the data based on their mutual similarity and without satisfactory descriptions provided by metadata self-contained in the documents and other content used.

Clustering can be performed on every kind of data: textual, visual and audio data as well as combination of these three. It is used especially for huge amount of data that are not well structured, or that are not structured at all. If it is not the case, some filtering and sorting are enough for preparing the data for information retrieval. Unfortunately, actual information systems are congested with different kind of content due to
long time of data accumulation, their distributed nature, demands to exchange the data with the other systems, different types of data and various formats that the data are stored in. Therefore, the software developers faced a complex problem how to integrate the same system functions to be applicable on such heterogeneous content. One of the most important pieces in the solution of this problem is clustering.

Basically, clustering represents a process of grouping data by using some algorithm or mathematical function. In both cases the calculating of similarity between data represents the main principle.

The considerations in the chapter are mainly related to clustering of textual content. In e-government IS the data are commonly stored in databases while the documents can be held in both the databases (DB) and repositories. Generally, DB provides easier way for grouping data and retrieving the information. There are eight parts in chapter. After the background briefly presented, the basic concepts used in clustering are described. Further, the common measures such as text frequency and inverse document frequency commonly used in clustering are described there. Moreover, some modifications of them as well as their combination are explained. The third section is about the clustering taxonomies. Many of them could be found in the research papers and the most common approach is followed—hierarchical and partition clustering represents the basic classification. Another one is also important: ‘hard’ (discrete) and ‘soft’ (fuzzy) clustering. For clarity the considerations are richly illustrated with the examples. In the fourth section the clustering techniques and algorithms are described. Two important techniques are presented: K-means and Fuzzy C-means. The fifth section is about different formats and structures used for representing text content. The case study about clustering in ADVANSE system is presented after. Finally, the future plans and conclusions are presented in the last two sections.

**Background**

In contemporary e-government IS the clustering represents one of the techniques for improving advanced search and information retrieval, providing of high quality of citizen services as a final goal. Further, such a goal cannot be achieved without clustering due to rapid growth of content in amount and number. On the other hand there are more citizens’ demands in form of questions and searching queries due to expansion of computer literacy and awareness of existence of E-Government services on the Internet (Fang, 2002). Therefore, the institutions need to improve their services in way to reduce the responding time and to provide higher quality of response.

The clustering exists more than 50 years (Krippendorff, 1980) but its development is strongly related with the progress in IT. Firstly it is used in anthropology (Driver & Kroeber, 1932) and psychology (Zubin, 1938) for building typologies of cultures and individuals by using empirical data. Since the early ‘60 its usage has been spread out from the human researches to biology, geography, astronomy, finance and so far. From the ‘just for researching’ analysis it has been developed to extraordinary useful set of techniques and algorithms that are implemented as modules of software products or as separate tools designed for business purposes. When the first book about clustering was published (Sneath & Sokal, 1973) there were numbers of applications as well as techniques and algorithms already built. Common for all of them is that from the beginning, clustering or cluster analysis has been dedicated.