Chapter XXII
Applied Sequence Clustering Techniques for Process Mining

Diogo R. Ferreira
IST – Technical University of Lisbon, Portugal

ABSTRACT

This chapter introduces the principles of sequence clustering and presents two case studies where the technique is used to discover behavioral patterns in event logs. In the first case study, the goal is to understand the way members of a software team perform their daily work, and the application of sequence clustering reveals a set of behavioral patterns that are related to some of the main processes being carried out by that team. In the second case study, the goal is to analyze the event history recorded in a technical support database in order to determine whether the recorded behavior complies with a predefined issue handling process. In this case, the application of sequence clustering confirms that all behavioral patterns share a common trend that resembles the original process. Throughout the chapter, special attention is given to the need for data preprocessing in order to obtain results that provide insight into the typical behavior of business processes.

1. INTRODUCTION

The field of process mining (van der Aalst & Weijters, 2004) is a new and exciting area of research, whose purpose is to develop techniques to gain insight into business processes based on the behavior recorded in event logs. There are a number of process mining techniques already available and most of them focus on discovering control-flow models (van der Aalst et al, 2003). There are also techniques that take into account data dependencies (Rozinat et al, 2006), and

techniques to discover other kinds of models such as social networks among workflow participants (van der Aalst et al, 2005).

Process mining techniques such as the α-algorithm (van der Aalst et al, 2004), the inference methods proposed by (Cook & Wolf, 1995), the directed acyclic graphs of (Agrawal et al, 1998), the inductive workflow acquisition by (Herbst & Karagiannis, 1998), the hierarchical clustering of (Greco et al, 2005), the genetic algorithms of (Alves de Medeiros et al, 2007) and the instance graphs of (van Dongen & van der Aalst, 2004), to cite only a few, are all techniques that aim at extracting the control-flow behavior of a business process and representing it according to different kinds of models. All of these techniques take an event log as input and as the starting point for the discovery of underlying process.

In many practical applications, however, the events that belong to a particular process can only be found among the events of other processes that are running within the same system. For example, events recorded in a CRM (Customer Relationship Management) system may belong to different processes such as creating a new customer or handling a claim submitted by an existing customer. Furthermore, even when focusing on a single process, the behavior in set of instances may be so diverse that it becomes appropriate to study different behaviors as separate workflows. Either way, the amount and diversity of activities recorded in an event log may be such that it becomes necessary to sort out the different existing processes before applying one of the above process mining techniques.

Sequence clustering is a particularly useful technique for this purpose, as it provides the means to partition a number of sequences into a set of clusters or groups of similar sequences. Although the development of sequence clustering techniques has been an active field of research especially in the area of bioinformatics—see for example (Enright et al, 2002), (Jaroszewski & Godzik, 2002) and (Chen et al, 2006)—its principles are equally applicable to other kinds of sequence data. For example, in applications such as user click-stream analysis it is possible to use sequence clustering to discover the typical navigation patterns on a Web site (Cadez et al, 2003). The same approach can be used to discover the typical behavior of different processes, or to distinguish between different behaviors within a single process, for example to identify what is considered to be the normal flow and what is deemed to be exceptional behavior.

The use of clustering algorithms in association with process mining techniques has received increased attention in recent years: in (Greco et al, 2004), the authors represent each trace in a vectorial space in order to make use of the *k*-means algorithm to cluster workflow traces; (Alves de Medeiros et al, 2008) make use of a similar approach in order to perform hierarchical clustering; (Jung et al, 2008) also address hierarchical clustering by means of a special-purpose algorithm based on a cosine similarity measure; in (Song et al, 2008) the authors make use of several clustering algorithms, including *k*-means and self-organizing maps; (Ceglowski et al, 2005) make use of self-organizing maps in order to cluster hospital emergency data. This means that there are several techniques available for clustering workflow traces. In this chapter we focus specifically on the use of sequence clustering techniques.

The chapter is organized as follows: Section 2 explains how sequence clustering works in order to find a set of clusters of similar sequences. Section 3 provides a word of caution regarding the need for preprocessing before actually applying sequence clustering to a given dataset. Section 4 presents a case study on the application of sequence clustering to an activity log that has been collected manually during the daily work of a software development team. Section 5 presents a second case study on the application of sequence clustering to the history recorded in a technical support system, in order to determine