Chapter IV
On the Usage of Structural Information in Constrained Semi-Supervised Clustering of XML Documents

Eduardo Bezerra
CEFET/RJ, Federal Center of Technological Education CSF, Brazil

Geraldo Xexéo
Programa de Sistemas, COPPE, UFRJ, Institute of Mathematics, UFRJ, Brazil

Marta Mattoso
Programa de Sistemas, COPPE/UFRJ, Brazil

ABSTRACT

In this chapter, we consider the problem of constrained clustering of documents. We focus on documents that present some form of structural information, in which prior knowledge is provided. Such structured data can guide the algorithm to a better clustering model. We consider the existence of a particular form of information to be clustered: textual documents that present a logical structure represented in XML format. Based on this consideration, we present algorithms that take advantage of XML metadata (structural information), thus improving the quality of the generated clustering models. This chapter also addresses the problem of inconsistent constraints and defines algorithms that eliminate inconsistencies, also based on the existence of structural information associated to the XML document collection.

INTRODUCTION

The problem of semisupervised clustering (SSC) has been attracting a lot of attention in the research community. This problem can be stated as follows: given a set of objects X and some prior knowledge about these objects, the clustering algorithm must produce a partition of X guided by this prior

knowledge. According to Grira, Crucianu, and Boujemaa (2004), there are two approaches for semisupervised clustering: *distance-based* and *constraint-based*. In distance-based semisupervised clustering, the prior knowledge about the data is used to modify the distance metric or the objective function in order to make distant objects farther and to make close objects closer (Chang & Yeung, 2004; Xing, Ng, Jordan, & Russell, 2002). In constraint-based semisupervised clustering, the prior knowledge is used to guide the clustering algorithm to a solution that reflects the user needs; this prior knowledge is usually in the form of *must-link constraints* and *cannot-link constraints* defined on the objects to be clustered (Basu, Banerjee, & Mooney, 2002, 2004; Wagstaff & Cardie, 2000). A must-link constraint $ML(o_i, o_j)$ states that objects o_i and o_j must be in the same cluster, whereas a cannot-link constraint $CL(o_i, o_j)$ states that o_i and o_j must be put in separate clusters. There are also hybrid approaches, which try both to learn a metric and to force the algorithm to obey the user-provided constraints (Basu, Bilenko, & Mooney, 2004; Bilenko, Basu, & Mooney, 2004).

Most semisupervised clustering algorithms are extensions of the well-known K-Means partitional clustering algorithm (MacQueen, 1967), although there are also approaches for hierarchical algorithms (Davidson & Ravi, 2005b). Experimental results show that the quality of the clustering models produced by these algorithms increases with the amount of provided prior knowledge. Nevertheless, despite the huge success of the semisupervised approach for clustering in recent years, there are still some open problems, especially when it comes to clustering of semistructured documents. Below, we summarize some of these problems.

- Associated to the characteristic of using external information, there is a first problem with current semisupervised clustering algorithms: they assume that the user is supposed to provide a significant amount of prior knowledge to allow the algorithm to produce a clustering model of a reasonable quality. However, in complex application domains (like textual document clustering), the user has to provide an amount of constraints that reaches the hundreds. Certainly, this is not a practical scenario. Usually, the user does not want (or is not able) to provide such a large amount of prior knowledge, particularly in an online system. Therefore, a first open problem in semisupervised clustering is to define approaches to reduce the amount of necessary constraints to be provided by the user.
- Another issue that we identified in current semisupervised clustering algorithms is that they are not prepared to take advantage of metadata. Thus, they require the provided constraints to be in the form $ML(o_i, o_j)$ or $CL(o_i, o_j)$, where o_i and o_j are two objects in the collection to be clustered. However, assuming that there is structural information associated with the collection of objects, the constraints could also be defined at such a level of abstraction. In other words, it should be possible for the user to define constraints in which the component objects are extracted from the metadata associated with the collection.

In addition, the amount of document collections associated to some form of structural information, particularly in XML (Bray, Paoli, & Sperberg-McQueen, 2000), is growing at a fast rate. The idea of the *Semantic Web*, for example, in which the Internet can be automatically navigated by software agents, assumes that data in this environment are in XML format. Another domain where textual documents are increasingly growing is in bioinformatics. Huge databases of textual information about proteins and amino acids, along with repositories of technical articles, have annotations or metadata information