Chapter IX
Featureless Data Clustering

Wilson Wong
University of Western Australia, Australia

Wei Liu
University of Western Australia, Australia

Mohammed Bennamoun
University of Western Australia, Australia

ABSTRACT

Feature-based semantic measurements have played a dominant role in conventional data clustering algorithms for many existing applications. However, the applicability of existing data clustering approaches to a wider range of applications is limited due to issues such as complexity involved in semantic computation, long pre-processing time required for feature preparation, and poor extensibility of semantic measurement due to non-incremental feature source. This chapter first summarises the many commonly used clustering algorithms and feature-based semantic measurements, and then highlights the shortcomings to make way for the proposal of an adaptive clustering approach based on featureless semantic measurements. The chapter concludes with experiments demonstrating the performance and wide applicability of the proposed clustering approach.

INTRODUCTION

Data clustering has a wide range of applicability ranging from ontology learning (Wong et al., 2006) and market research, to pattern recognition and image processing (Jain et al., 1999). Depending on the areas of applications, many different names such as cluster analysis, numerical taxonomy, automatic classification, botryology and typological analysis have been devised to refer to essentially the same practice of data clustering. Each application of data clustering can be characterised by two aspects, namely,
the manner through which the clusters are formed, and the criteria that govern the formation of clusters. The first aspect relates to the choice of clustering algorithm while the second aspect dictates the type of semantic measure (i.e. similarity or relatedness) to be used. The choice of the clustering algorithms, and semantic measures very much depends on the data elements to be clustered, computational constraints, and also the desired results. In the past, data clustering have been particularly successful with certain types of data such as documents, software systems, lexical units, webpages, Uniform Resource Identifiers (URIs) and images. This gives rise to various sub-areas of clustering such as document clustering (Steinbach et al., 2000), software botryology (Tzerpos and Holt, 1998), term clustering (Wong et al., 2007), webpage clustering (Wang and Kitsuregawa, 2002), usage-based URI clustering (Mobasher et al., 1999) and clustering-based image segmentation (Jain et al. 1999).

Unlike documents, webpages, software systems, and images or visual objects, terms and URIs are essentially featureless. The semantic relation between visual objects can be established by feature analysis based on visible (e.g. physical and behavioral) traits. The metadata and text that documents and webpages contain can be useful as features. However, establishing the semantic relation between data elements such as terms depends on something less tangible, namely, background knowledge which humans acquired through their senses over the years. Similarly, URIs alone are essentially strings of characters whose significance can be difficult to establish without visiting the actual resources such as webpages. Along the same line of thought, Lagus et al. (1996) state “In principle a document might be encoded as a histogram of its words... symbolic words as such retain no information of their relatedness”. To illustrate, consider how do one know that a cow has four legs without actually looking at it or acquiring such fact from someone else? Equally challenging is how can one establish that the URI http://del.icio.us/ actually refers to a social bookmark site and not a website related to cooking or food without prior exposure? In this chapter, we refer to documents, webpages, images and software systems as decomposable data while non-decomposable data is used exclusively to denote terms and URIs. Decomposable data elements are those that can be meaningfully analysed in terms of their components or constituent parts for uncovering semantics or significance. Obviously, having the ability to take apart a data element for extracting features to assist in further stages of semantic relation analysis can be very helpful in data clustering. Putting aside the computational complexity and complications related to feature extraction and feature selection, clustering involving decomposable data can be much more objective since the criteria for cluster formation can be explicitly identified.

The focus of this chapter is on the clustering of non-decomposable data (i.e. terms and URIs) where the absence of explicit features poses certain challenges, which requires adjustments to be made, especially with regard to the computation of semantic relation, and the clustering algorithm itself. Firstly, a recent survey in ontology learning (Gomez-Perez and Manzano-Macho, 2003) reveals that all reviewed systems which apply term clustering methods rely on contextual cues surrounding the terms as features. However, the large collection of documents, and predefined patterns and templates that are required for the extraction of contextual cues makes the portability and extensibility of such systems difficult. Similarly, the conventional approach to URI clustering for Web personalization requires large amount of log records to enable the use of feature-based semantic measures. Consequently, featureless semantic measures are fast becoming a necessity for clustering non-decomposable data. Secondly, to compensate for the lack of configurable and objective measurement of semantic relation through the use of features, changes to the clustering algorithm is also necessary. In feature-based semantic measurements, the type and granularity of features can be selected or refined depending on application requirements. For example, the choice of features for face recognition can come from certain regions of the face such