Chapter 2

Preliminaries

This chapter addresses the basic concepts required for efficient similarity search in non-vector databases, such as image and video databases. First, Section 2.1 presents an overview of feature representation models including histograms and signatures. After that, Section 2.2 focuses on query types which are often used in similarity search. Section 2.3 is devoted to the prominent distance-based similarity measure Earth Mover's Distance which is utilized throughout this thesis. This chapter is concluded by Section 2.4 which gives an overview about efficient similarity search.

2.1 Feature Representation

In order to process and store data arising in many application domains, the data needs to be represented and stored mathematically for which appropriate feature representation models are required. One of the most encountered approaches is to represent each single data record by a feature vector in a feature space. Take the restaurant locations in Los Angeles as an example: each restaurant is represented by its coordinates in a 2-dimensional vector space. While vector data model is a simple representation model, nevertheless there are various kinds of data for which vector representation model is not mathematically adequate. Figure 2.1 illustrates an example from an online movie store where customer Q can be identified by the number of DVD's

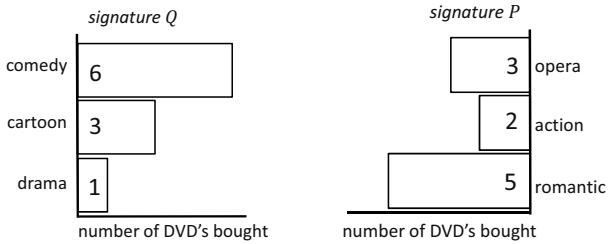

Figure 2.1: Feature representations of two customers in an online movie store. Each customer is identified by a particular number of representatives with individual weights.

bought in categories *comedy, cartoon* and *drama*, while it is observed that customer P bought items in categories *opera, action*, and *romantic*. Categories in which DVD's were bought help those customers to be distinguished from each other easily. In addition, the number of DVD's depicted in each bucket is another mark which additionally contributes to the representation of each customer. As illustrated in this example, features which are appropriate and important for the representation of data objects may vary from each other which facilitates differentiation. Such feature representations are called *signatures* which are variable-length distributions over specific feature vectors in the underlying feature space. In other words, each signature may exhibit an individual number of feature vectors, as well as possibly different feature vectors with possibly various number of feature vectors assigned to them in the underlying feature space. The number of feature vectors assigned to the corresponding feature vector is often described as *weight* of that feature vector. Feature vectors which exhibit weights greater than zero are denoted as *representatives* in order to differentiate them from those feature vectors which have weights equal to zero. For instance, signature Q in the example above has three representatives, namely *comedy, cartoon* and *drama* with the weights 6, 3, and 1, respectively. Any other feature vectors located in the feature space do not contribute to the representation of this