# 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

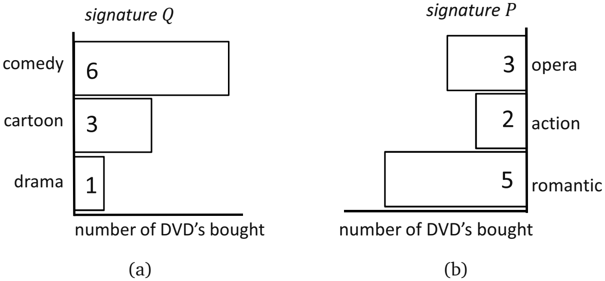
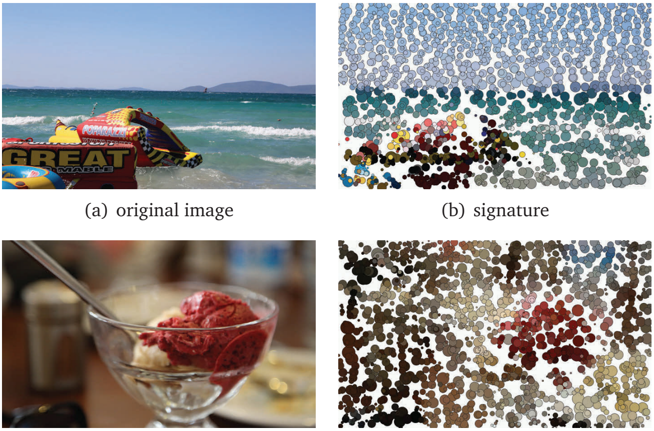


Figure 2.1: Feature representations of two customers in an online movie store. Each customer is identified by a particular number of representatives with individual weights.

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



(a) original image

(b) signature

(c) original image

(d) signature

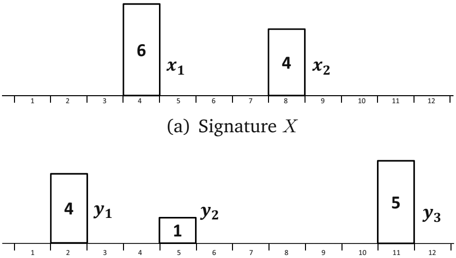
Figure 2.2: Example images and the visualization of their signatures with 1000 representatives.

customer and, hence, do not belong to the representative set of signature $Q$. The formal definition of signature is given as follows.

**Definition 2.1 (signature)** *Given a feature space* $\mathbb{F}$ *and a ground distance function* $\delta$*, a* signature $X : \mathbb{F} \to \mathbb{R}$ *is defined as a mapping from the feature space to the real numbers for which the representative set* $R_X$ *is finite, i.e.* $|R_X| < \infty$ *holds.*

Figure 2.2 illustrates two images[*] and their corresponding signatures as examples from multimedia domain [UBS15]. Figure 2.2(b) and (d) exhibit signatures whose representatives are feature vectors found in a 5-dimensional feature space with 2 dimensions of location information (horizontal and vertical location) and 3 color dimensions (CIE Lab). Here, each image is represented by 1000 representatives with some weights. In signature visualizations, the center of each circle denotes a representative and the radius of

[*] The pictures are taken by the author.



(a) Signature $X$

(b) Signature $Y$

Figure 2.3: Illustration of two signatures $X, Y$ whose representatives are visualized in a 1-dimensional feature space.

each circle corresponds to the weight of that representative. One possible and common way to acquire representatives and their weights is to cluster feature vectors of the data object by an appropriate clustering algorithm, such as k-means. The resulting cluster centroids correspond to representatives and the number of feature vectors assigned to a single representative denotes the (absolute) weight of that representative.

Figure 2.3 depicts two signatures $X$ and $Y$ with representative sets $R_X = \{x_1, x_2\}$ and $R_Y = \{y_1, y_2, y_3\}$ which are represented in a 1-dimensional feature space. The indices of feature vectors are presented by the numbers between 1 and 12 for each signature. Since feature vectors exhibiting a weight equal to zero do not correspond to the representation of the underlying data objects, they are not illustrated in the figure, instead, only representatives are highlighted by their weights in the buckets: $X(x_1) = 6$, $X(x_2) = 4$ and $Y(y_1) = 4$, $Y(y_2) = 1$, $Y(y_3) = 5$.

Many domains and applications require data objects to be represented by feature distributions. Example data types include biological data[HL03, NCA$^+$13], multimedia data[RTG98, UBSS14], medical data[CH15], uncertain data[CCCX09], probabilistic data [XZTY10], telecommunication data [ADKU11], gestures [RYZ11], and text documents [BGD15].

In order to save storage and increase efficiency of query processing, many applications utilize *a priori* partitioning of the underlying feature space into fixed-size partitions (a.k.a. bins). In this way, the number of feature vectors belonging to its partition determines the weight of that corresponding bin, leading to a *histogram* representation. Since bins remain unchanged for all data objects, comparing two histograms with each other is then only based on taking weights of bins into consideration. As a natural result, query processing time is decreased and storage requirements are reduced.

The representation of data objects by histograms may result in reduced expressiveness of those data objects, if compared with signatures [RTG98]. Take the feature representation $Y$ in Figure 2.3 as an example. If $Y$ were represented by using an *a priori* partitioning of the feature space where only the representatives $x_1$ and $x_2$ were utilized as bins, the histogram representation of $Y$ would be determined as $(0, 0)$, since $Y$ exhibits the weight of zero for both bins. Hence, such a histogram representation would not lead to an expressive representation of the corresponding data object. As elucidated through this example, expressiveness can be limited for data objects if an *a priori* partitioning of the underlying feature space is used. Since there is a trade-off between expressiveness of data object representation and efficiency of query processing based on that, it depends on priorities and requirements of the user and system, which feature representation model should be used. In this thesis, due to better expressiveness of data objects via flexible representation of feature vectors by some weights in the underlying feature space, we will focus on efficient query processing approaches based on the signature representation model. The signature representation model enables each feature vector in the feature space to exhibit a weight (which is assumed to be non-negative in this work), and each representative is assigned a positive real number denoting the weight of that representative. Theoretically, there exist infinitely many feature vectors which are used to represent the corresponding data object, but since mathematically a vector can be represented by finitely many dimensions, we take those feature vectors into consideration which exhibit a non-zero weight. Initially introduced by Rubner *et al.* in [RTG98], signatures were also investigated in [Bee13], as well. Note that the terms *histogram* and *signature* are used to denote *feature histogram* and *feature signature*, respectively, and both variations can be found in the literature