Data Management in Three-Dimensional Structures

Xiong Wang
California State University at Fullerton, USA

INTRODUCTION

Data management in its general term refers to activities that involve the acquisition, storage, and retrieval of data. Traditionally, information retrieval is facilitated through queries, such as exact search, nearest neighbor search, range search, etc. In the last decade, data mining has emerged as one of the most dynamic fields in the frontier of data management. Data mining refers to the process of extracting useful knowledge from the data. Popular data mining techniques include association rule discovery, frequent pattern discovery, classification, and clustering. In this chapter, we discuss data management in a specific type of data i.e., three-dimensional structures. While research on text and multimedia data management has attracted considerable attention and substantial progress has been made, data management in three-dimensional structures is still in its infancy (Castelli & Bergman, 2001; Paquet & Rioux, 1999). Data management in 3D structures raises several interesting problems:

1. Similarity search
2. Pattern discovery
3. Classification
4. Clustering

Given a database of 3D structures and a query 3D structure, similarity search looks for those structures in the database that match the query structure within a range of tolerable errors. The similarity could be defined in two different measurements. The first measurement compares the data structure with the query structure in their entirety i.e., a point-to-point match. We will call this aggregate similarity search. The second measurement compares only the contours or shapes of the data structure with that of the query structure. This is generally referred to as shape-based similarity search. The range of tolerable errors specifies how close the match should be when the data structure is aligned with the query structure. Pattern discovery is concerned with similar substructures that occur in multiple structures. Classification and clustering when applied to these domains attempt to group 3D structures with similar shapes or containing similar patterns together.

BACKGROUND

Three-dimensional structures can be used to describe data in different domains. In biology and chemistry, for example, a molecule is represented as a 3D structure with connections. Each point is the center of an atom and the connections are bonds between atoms. In computer-aided design, an object is specified as a set of 3D vectors that describes the shape of the object (Veltkamp, 2001; Suzuki & Sugimoto, 2003). In computer vision, the shape of a 3D object can be caught by X-ray or ultrasonic scanning devices. The result is a set of 3D points (Hilaga, Shinagawa, Kohmura, & Kunii, 2001). In medical imaging, 3D images of tissues or tumors can be collected using magnetic resonance imaging or computer tomography (Akutsu, Arakawa, & Murase, 2002). With advances in the Internet, scanning devices, and storage, the World Wide Web is becoming a huge reservoir of all kinds of data. The 3D models available over the Internet dramatically increased in the last two decades. Similarity search is a highly desirable technique in all these domains. Classification and clustering of biological data or chemical compounds have special significances. For example, traditionally proteins are classified to families according to their specific functions. However, recently, many approaches have been proposed to classify proteins according to their structures. Some of these approaches achieve very high accuracy when compared with their biological counterparts. Classification and clustering can also help build index structures in 3D model retrieval to speed up similarity search.

Currently, there is not a universal model or framework for the representation, storage, and retrieval of three-dimensional structures. Most of these data are stored in plain text files in some specific format. The format is different from application to application. Likewise, the existing techniques for information retrieval and data mining in three-dimensional structures take root in the areas of application. Two main areas of application are computer vision and scientific data mining, where computation-intensive techniques have been developed and are still in demand. We focus on data management in these two areas.

ADVANCES IN COMPUTER VISION

Shape-based recognition of 3D objects is a core problem in computer vision and has been studied for decades. There are roughly three categories of approaches: volume-based, feature-based, and interactive. For example, Keim (1999) proposed a geometric-based similarity search tree to deal with 3D volume-based search. He suggested using voxels to approximate the 3D objects. For each 3D object, a Minimum Surrounding Volume (MSV) and Maximum Including Volume (MIV) are constructed using voxels. Similar objects are clustered in data pages and the MSV and MIV approximations of the objects are stored in the directory pages. Another interesting scheme uses superellipsoids to approximate the shape of the volume. Superellipsoids are similar to ellipsoids except the terms in the definition are raised to parameterized exponents which are not necessarily integers. Indexing the superellipsoids is very difficult due to their different shapes and sizes.

Feature-based approaches have been developed extensively in the literature. In (Belongie, Malik, & Puzicha, 2001, 2002), Belongie and co-authors designed a descriptor, called the shape context, to characterize the distribution of the structure. For each point p_i in the structure, the set of 3D vectors originating from p_i to every other point in the structure is collected as the shape context. A histogram is constructed based on comparison between the shape context of the query shape and that of the data shape. Osada, Funkhouser, Chazelle and Dobkin (2001) suggested using a similar descriptor, called the shape distribution. The shape distribution is a set of values that are calculated by a shape function, such as the Euclidean distance between two randomly selected points on the surface of a 3D object. A histogram is calculated based on the shape distributions of the two shapes under consideration. Kriegel *et al.* introduced a 3D shape similarity model that decomposes the 3D model into concentric shells and sectors around the center point (Ankerst, Kastenmüller, Kriegel, & Seidl, 1999). The histograms are determined by counting the number of points within each cell. The similarity is calculated using a quadratic form distance function. Korn, Sidiropoulos, Faloutsos, Siegel and Protopapas (1998) proposed another descriptor, called size distribution that is similar to the pattern spectrum. They introduced a multi-scale distance function based on mathematical morphology. The distance function is integrated to the GEMINI framework to prune the search space. GEMINI is an index structure for high dimensional feature spaces. Bespalov, Shokoufandeh, Regli, & Sun (2003) used Singular Value Decomposition to find suitable feature vectors. A data level shape descriptor, called the spin image, was used in (Johnson & Hebert, 1999) to match surfaces represented as surface meshes. The system is capable of recognizing multiple objects in 3D scenes that contain clutter and occlusion. Saupe and Vranic (2001) suggested using rays that cast from the center of the mass of the object as feature vectors. The representation is compared using spherical harmonics and moments. All these descriptors are very high dimensional spaces and indexing them is a well known difficult problem, due to "the curse of dimensionality". Furthermore, the approaches are not suitable for pattern discovery.

An interactive search scheme was introduced in (Elad, Tal, & Ar, 2001). The algorithm sets the center of a structure to the origin and calculates the normalized moment value of each point. The normalized moment values are used to approximate the structure. Two structures are compared according to a weighted Euclidean distance. The weights are adapted based on the feedback from the user, using Singular Value Decomposition. Chui and Rangarajan (2000) developed an iterative optimization algorithm to estimate point correspondences and image transformations, such as affine or thin plate splines. The cost function is the sum of Euclidean distances between the transformed query shape and the transformed data shape. The distance function needs an initial correspondence between the two structures. Interactive refinement was also used in a medical image database system developed by Lehmann *et al.* (2004). A survey of shape matching in computer vision can be found in (Veltkamp & Hagedoorn, 1999).

Shape-based similarity search in a large database of 3D objects is much more challenging and is still a young research area. The most recent achievements are a search engine developed at Princeton University (Funkhouser, *et al.,* 2003) and a search system 3DESS for 3D engineering shapes built at Purdue University (Lou, Prabhakar, & Ramani, 2004). These techniques were concentrated on computer vision and they did not compare the 3D models point-to-point. Many of them only search for 3D objects that look similar. The effectiveness of such techniques often depends on subjective perception.

EMERGING TECHNIQUES IN SCIENTIFIC DATA MINING

In structural biology, detecting similarity in protein structures has been a useful tool for discovering evolutionary relationships, analyzing functional similarity, and predicting protein structures. The differences in proteins or chemical compounds are very subtle. To discover functionally related proteins or chemical compounds, in many cases we have to match them atom-to-atom. Aggregate similarity search is known as an *NP* complete problem in combinatorial pattern matching. Even though the significance of the problem surfaced with applications to