Chapter 8

G-Hash: Towards Fast Kernel-Based Similarity Search in Large Graph Databases

Xiaohong Wang
University of Kansas, USA

Jun Huan
University of Kansas, USA

Aaron Smalter
University of Kansas, USA

Gerald H. Lushington
University of Kansas, USA

ABSTRACT

Structured data such as graphs and networks have posed significant challenges to fundamental aspects of data management including efficient storage, indexing, and similarity search. With the fast accumulation of graph databases, similarity search in graph databases has emerged as an important research topic. Graph similarity search has applications in a wide range of domains including chemoinformatics, bioinformatics, sensor network management, social network management, and XML documents, among others.

Our objective in this chapter is to enable fast similarity search in large graph databases with graph kernel functions. In particular, we propose (i) a novel kernel-based similarity measurement and (ii) an efficient indexing structure for graph data management. In our method, we use a hash table to support efficient storage and fast search of the extracted local features from graph data. Using the hash table, we have developed a graph kernel function to capture the intrinsic similarity of graphs and for fast similarity query processing. We have demonstrated the utility of the proposed methods using large chemical structure graph databases.

DOI: 10.4018/978-1-61350-053-8.ch008
INTRODUCTION

Structured data including sets, sequences, trees and graphs, pose significant challenges to fundamental aspects of data management such as efficient storage, indexing, component search (e.g. subgraph/supergraph search) and similarity search. Among these structured data, the use of graph representations has gained popularity in pattern recognition and machine learning. The main advantage of graphs is that they offer a powerful way to represent structured data. Querying and mining of the graphs can contribute to our understanding in numerous ways: understanding of new connectivity patterns, evolutionary changes and discovery of topological features. Queries in graph databases can be broadly classified into two categories: (i) subgraph query and (ii) similarity query. The aim of subgraph query is to identify a set of graphs that contain a query graph. The aim of similarity query is to identify similar graphs in a graph database to a query, according to a distance metric. There are two types of similarity query, i.e. k-NNs query and range query. In k-NNs query, the k most similar graphs are reported. In range query, all graphs within a predefined distance to the query graph are reported. In this chapter, we address the problem of k-NNs similarity search in large database of graphs.

Motivation

Recent scientific and technological advances have resulted in an abundance of data modeled as graphs. Graph has been used extensively in modeling complicated structures and schemaless data, such as proteins (Berman et al., 2000), image (Berrett et al., 2001), visions (Fu, 1986), program flows (Liu et al., 2005), XML documents (McHugh et al., 1997), the Web (Faloutsos et al., 1999), etc. For example, a metabolic pathway is modeled as a set of reactions, enzymes and metabolites, and an edge is placed between a reaction and a metabolite (or enzyme) if it participates in the reaction and schema of heterogeneous web-based data sources and e-commerce sites can also be modeled as graphs (He and Singh, 2006). With the fast growth of graph databases, similarity search in graph databases has emerged as an important research topic. Graph similarity search has applications in a wide range of domains including chemoinformatics, bioinformatics, sensor network management, social network management, and XML documents, among others. For example, in chemistry and pharmacy, there is a fast accumulation of chemical molecule data. Once a new chemical is synthesized, the properties of the chemical may be revealed through querying existing chemicals with known properties. Fast similarity search in large graph databases enables scientists and data engineers to build accurate models for graphs, identify the intrinsic connections between graph data, and reduce the computational cost of processing large databases.

Many researchers have been working on the graph similarity search. The most straightforward approach for similarity measurement is to embed a graph in a high dimensional Euclidian space, known as the feature space, uses spatial indexing techniques for similarity search. Much of this work has naturally tended toward methods that balance the expressivity of substructures with their efficiency of enumeration. General subgraphs are the most expressive, as they are graphs themselves, but that also makes them the most expensive enumeration. The application of frequent subgraph mining has seen much success in this area, as a way to retain high expressivity while allowing fast enumeration. At the other end of the spectrum, sets or bags of vertices and edges are the least expressive and ignore all of the rich information of graph connectivity; hence set representations for graphs are not widely used. Falling between these two extremes are intermediate structures such as paths, cycles, and trees. Substructure enumerations are then the basis for many useful operations such as hashing, indexing, similarity search, etc. In all, current feature extraction meth-