Chapter 8

Entity Resolution on Graph Data Set

ABSTRACT

In this chapter, the authors study entity resolution on graph data set. In order to conduct entity resolution on graph data, the authors need to define the distance of graph. The authors compute these distances or approximately compute them for time efficiency. At last, the authors utilize the distances to get the final result of entity resolution. The approximate graph matching algorithms may be index-based like the NH-Index method or kernel function based like G-hash method. Other methods concentrate on providing new definitions of similar graph that are easier to compute than traditional methods, like the Web-collection method and the Grafil method. To increase the resolution ability of traditional methods, researchers provide some methods to recognize similar graphs, like graph-bounded simulation and p-homomorphism. Section 8.1 introduces existing methods on defining the distance of graph, which has a direct impact on the computation of graph similarity. Section 8.1 introduces pair-wise entity resolution on graph data set, including index techniques, graph-bounded simulation, and graph p-homomorphism.

INTRODUCTION

At earlier times, researchers concentrate on exact graph matching, for exact graph matching can get the same subgraph in the data graph. However, as the development of database theory, more none-relational database systems come into the world, and graph database system is one of the most suitable systems for the processing of big data. For many times, exact graph matching cannot express the query intention well. User queries include querying for whether there exists web store with similar production structure, querying for protein molecules with a given protein molecules, etc. As the Chinese saying goes, “no clear water to fish”, exact graph matching may lead to not enough number of matching results or even no matching result. So similar graph matching plays an important role in the database management world.

The approximate graph matching algorithms may be index-based like the NH-Index method, or kernel function based like G-hash method.

NH-Index(Tian, 2008), short for neighborhood index is an easy to implement index structure for similar graph matching. Traditional graph index methods only index subgraphs (paths, trees or general subgraphs), which can lead to index sizes that are exponential in the database size. The index...
unit for NH-Index is the neighbor information of each node in database and the index size is linear in the database size. Also, the NH-Index is a disk-based index, which is suitable for big data that cannot be all put in main memory.

G-hash (Wang, 2009) is a kernel function based method to do pair-wise entity resolution on graphs. The basic idea of this method is mapping graph data into node vectors, and we can get graph similarity by computing similarity function on these node vectors. In order to get node vector, we first make use of wavelet functions, which transform the topology of graphs into node vectors. Kernel function refers to the operation of computing the inner product between two objects in feature space. Kernel function computes the similarity of node vectors, which reflect the similarity between graphs.

Other methods concentrate on providing new definitions of similar graph that are easier to compute than traditional methods, like the web-collection (Cho, 2000) method and the Grafil(Yan, 2005) method.

Web collection is a practical graph similarity measure method defined by the group who developed the Google search engine. Their aim is to find replicated web pages, and this is done by modeling web pages as a web graph. The basic processing unit is called collection in this method.

Grafil is a similarity measure for graph. This measure builds a connection between the structure-based measure and the feature-based measure so that we can use the feature-based measure to screen the database before performing the expensive pairwise structure-based similarity computation. When performing subgraph matching, too strict matching will induce a nearly empty result set. Grafil’s basic idea is to extract features from query graph, and when the result set doesn’t have enough elements, we gradually reduce the number of features to return more similar results. This process is called query relaxation. During the computation, feature filtering can then improve time efficiency.

To increase the resolution ability of traditional methods, researchers provide some methods to recognize similar graphs, like graph bounded simulation (Fan, 2010) and p-homomorphism (Fan, 2010).

Informally speaking, the graph isomorphism problem is to judge whether one graph $G_1$ can turn into $G_2$ only by relabelling node id. The subgraph isomorphism problem is to determine whether $G_1$ is isomorphism to some subgraph of $G_2$. Subgraph isomorphism is a NP-complete problem and graph isomorphism problem is a NP-hard problem. The graph homomorphism problem is to find a mapping from $V_1$ to $V_2$ such that each node in $V_1$ is mapped to a node in $V_2$ with the same label, and each edge in $E_1$ is mapped to an edge in $E_2$, while the subgraph isomorphism problem is to find a 1-1 mapping.

We can modify the problems as follows. For node match, we introduce a similarity matrix $\text{mat()}$ to measure the similarity of two nodes $u$ and $v$ between pattern graph and data graph. The matrix can be computed in terms of common shingles that $u$ and $v$ share. Then we can extend the mapping process. We can map edges in pattern graph to paths in data graph. The $P$-homomorphism and 1-1 $P$-homomorphism are extensions of graph homomorphism and subgraph isomorphism, respectively.

Bounded simulation is an extension of graph simulation. Its biggest change to graph simulation is mapping edges in pattern graphs to paths in data graphs, and the path’s length is a user-controllable parameter $k$. We can see graph simulation as a special case of bound simulation when $k = 1$.

**BACKGROUND**

Defining the distance of graphs is an important problem for researchers. People have some successful definitions such like the graph edit distance and graph alignment distance. They are good to measure similarity but difficult to compute.
Related Content

Super Computer Heterogeneous Classifier Meta-Ensembles

Employing Neural Networks in Data Mining

A Multidimensional Pattern Based Approach for the Design of Data Marts

A Framework for Efficient Association Rule Mining in XML Data