CSE 705 – Seminar Report
Discovering Topical Structures of Databases

Professor: Michalis Petropoulos

Megha Ramesh Kumar (meghar@buffalo.edu)

Overview:

In today's enterprise world, the scale of the databases and the increasing complexity of these databases and the prevalent lack of documentation make it hard for a data architect to understand, reverse engineer and integrate the databases. In this paper, the problem of discovering topical structures of databases to support semantic browsing and large scale data integration is addressed. The iDisc approach, a novel multi-strategy discovery framework and a novel clustering aggregation technique is proposed in this paper. Starting with a formal definition of the paper, the iDisc approach is explained in detail with the various algorithms for each module.

iDisc consists of a four module system namely model builder, base clusterers, meta-clusterers and the representative finder. The model builder examines the database from a number of perspectives and obtains a variety of representations which are explained in the paper such as the vector-based representation which captures the topical structures via the descriptive information. Here each table is represented as a text document and database is a collection of documents. The structures of the individual tables are ignored. The graph-based representation captures the topical structure via the linkage among tables. The similarity-based representations capture the topical structure via the value based similarity between the tables. Next, the base clusterers take a database representation and discover preliminary clusterings over the tables in the database. It implements several generic clustering algorithms such as the similarity-based algorithm or the linkage-based algorithm and instantiates them into clusterers. For the vector based representations and similarity based representations, generic similarity-based algorithm is instantiated while for graph-based representations the generic linkage-based algorithm is instantiated. The similarity-based generic algorithm evaluates the quality and a (ClsrSim) cluster similarity function which can be implemented in way such as single-link, complete-link, and average-link. The cluster with the maximum clustering quality is returned. The linkage-based generic algorithm has an EdgeDel function that suggests the edges to be removed. Implementations of EdgeDel are Shortest-path betweenness and Spectral graph partitioning.

The meta-clusterer which is the third module in the iDisc framework finds a clustering such that it agrees with the preliminary clusterings from the base clusterers as much as possible. It aggregates those clusters to give the final clusters with a key condition that it tries to minimize the disagreements between the preliminary clusters. The last module which is the representative finder discovers important tables within each cluster. These tables which are called cluster representatives serve as entry points to the cluster and give the users a general idea of what the cluster is about.

Detailed Comments:

Strengths: The key strength is that the proposed framework is highly extensible where additional database representations and base clusterers can be easily incorporated to the system. This helps in improvisation of the performance of the system. It provides a principled way of combining evidence through aggregations of votes from the clusterers rather than directly combining the disparate evidence from the database.A key problem of primary keys and foreign keys missing in catalogs is implemented by discovering primary keys and then proceeding to discover the foreign keys. The experimental results given in the paper show the effectiveness of the proposed iDisc framework. Experiments over several large real-world databases indicate that iDisc is highly effective, with an accuracy rate of up to 87%.

Weakness: Not much information has been given about handling complex aggregations. The experiments were conducted in a controlled manner. More examples of databases and the performance of iDisc could have been compared with each other giving us an idea about the weakness about the proposed approach.

Technical Depth: This paper covers various algorithms which are in-depth and provides a well developed architecture and accuracy values for experimental results. Paper is very technically sound since it covers algorithms in depth and a novel strategy framework which is used to address the key issue that we face today of increasing complexity of databases and lack of their documentation and the problem of reverse engineering to integrating the databases.

Comparison with Related Work: Data modeling products allow users to organize entities in a large logical model by subject areas during a top-down modeling process, to cope with the complexity and facilitate a modular development. The solution provided in this paper complements these functions by enabling users to reverse-engineer subject areas from a large-scale physical database during a bottom-up modeling process.
Incremental Schema Matching proposes a fragment oriented approach to matching large schemas to reduce matching complexity. It decomposes large schemas into several fragments and performs fragment-wise matching. The solution proposed in this paper complements this proposal by providing an automatic approach to partitioning a large schema into semantically meaningful fragments.

Discussions: Some of the discussions were related to the algorithms employed by the base clusterers and the various representations generated by the model builder. There were also comments made on the empirical evaluations. Questions were raised about rules employed for the various representations developed by the model builder.