CSE 718 Seminar Report

A Relational Approach to Incrementally Extracting and Querying Structure in Unstructured Data

Presenter – Hua Zhao
hzhao5@buffalo.edu

Overview:

This paper explores using a relational system as the basis for a workbench for extracting and querying structure from unstructured data. It shows that the data set is always available for some form of querying, and that as it is processed, users can pose a richer set of structured queries.

The following are the three main contributions of this paper:

1. This relational workbench provides:
    - A way to store an expanding set of documents and attributes.
    - Tools to incrementally process the data.
    - A way to exploit structure in queries.
2. Applies the relational approach to support structured queries over Wikipedia.
3. Provides examples of how to incrementally evolve the understanding of the data in the context of the relational workbench.

The paper uses wide table with complex attributes and mapping table and relationship table to store the unstructured data.

This paper provides three basic operators and six possible pair wise combinations of these three distinct operators.
Extract Operator: To Extract structure from unstructured data. The output of an extract operator is a set of structures.
Integrate Operator: Take as input a set of structures from the wide table or a previous operator and returns one or more sets of mapping over attributes that correspond to the same real-world concept.
Cluster Operator: Take in a set of documents or a set of attributes and classifies the input into one or more clusters.
Six possible pair-wise combinations: Integrate-Extract

- Cluster-Extract
- Extract-Cluster
- Integrate-Cluster
- Extract-Integrate
- Cluster-Integrate

This paper presents four steps to construct an example workbench for Wikipedia:

Stage 1: Initial Loading

Parses and loads XML files to wide table.

Stage 2: Extracting SectionName(text)

From each page, extracts the structure SectionName(text), in which SectionName represents the name of a first-level section in the page and the text is the content in that section.

Stage 3: Extracting info box as a blob

Extracts info boxes, which are general templates that contain predefined attributes and vary depending on the domain.

Stage 4: Extracting structured data from info boxes and wiki tables

Extracted and queried structured data from info boxes and wiki tables. Transformed the contents of the temperature wiki table into a relation.

Detailed Comments

1. The main strength of this paper is that it proposes a relational workbench as a frame work for extracting and querying structure from unstructured data. Also to get to incrementally know of the unstructured data, this paper combines extract, integrate, cluster operators and applied repeatedly to keep evolving understanding of the data set. It also gives us a concrete example to show this frame work is feasible.
2. The weakness of this paper is that it does not represent novel approach or in-depth analysis regarding extracting and querying structure from unstructured data. This paper is more concerned about the integration of various existing ideas and technologies from the platform view rather than scientific research.
   Another weakness is that it does not compare with other workbench platforms implemented by other organizations.
3. This paper is well balanced in terms of technical depth. It begins by describing the motivation behind this paper and a clear introduction of the basis used for the relational workbench. For the operator part, it gives brief and clear explanation, although it has far more details. But they are not the subject of this paper.
4. The paper is technically sound. The approach of relational workbench is straightforward. Moreover, the authors implement this workbench over Wikipedia.
5. There is a large body of literature relevant to various aspects of the workbench model, none of which employs their approach inside a relational database in an end-to-end fashion. For most of them, the use of a relational database is limited to storing a set of structures, which usually have a well-defined schema when they are loaded into the database. In contrast, this paper focuses on the idea of incrementally processing data in a relational database.