IntelliLIGHT: A Flashlight for Large-scale Dark Structured Data

Michael Gubanov, Manju Priya, Maksim Podkorytov
Department of Computer Science
University of Texas at San Antonio

I. Introduction

Large enterprises are continuously looking for new data sources to enrich their warehouse data. By leveraging more data, they can obtain more accurate product sales predictions and restock accordingly for the next year. For example, enterprise analysts believe that having a complete weather data would greatly improve prediction accuracy of beer sales in a region, hence enriching their existing warehouse data with weather data would be invaluable. However, their previous experience dealing with large-scale data sources suggests that data acquisition costs are high. As a result, they never enrich their data with complete weather dataset, hence cannot benefit from it in their predictions. `IntelliLIGHT` is a new system that quickly sheds light into the contents of a new and unknown, large-scale structured dataset, to help locate and retrieve needed data.

Figure II illustrates a screenshot of `IntelliLIGHT` interface displaying a summary of an imported large-scale structured dataset (*36 million* cleaned Web tables). The left pane of the interface has a treeview of major objects in the dataset, ranked and clustered by *ObjectRank* - a new ranking function for large-scale structured data. *ObjectRank* helps identify the most significant tables among millions available. Similar to PageRank [Brin and Page, 1998] that operates on the web graph, we construct a large graph from the pertinent tables in the dataset and account for the direction and degree of a node in this graph to calculate the node rank. Because of ranking, users can quickly understand what are the main contents of a large-scale dataset, without having to fully explore it. Without *ObjectRank*, the user is faced with millions of tables and a problem of finding needed information among them. There would be no summary, so even if the user can see a sample of the dataset, it is completely unclear what does the remainder contain. Also, there would be no no way to rank tables, hence no way to understand what objects have the most footprint in the dataset. In fact, it would be similar to a Web search engine without a ranking function.

To locate and extract main tables among millions with high accuracy, we use large-scale machine learning algorithms that we developed. It is extremely difficult if not impossible to accomplish this task both with high accuracy and at scale, without machine learning. Any machine learning algorithm has a training phase, which usually requires manual effort, labeled training data, hence is expensive and time consuming, which is prohibitive at scale. Our second contribution is an algorithm to reduce manual effort by generating object-identifying classifiers automatically based on a few descriptive keywords from the user.

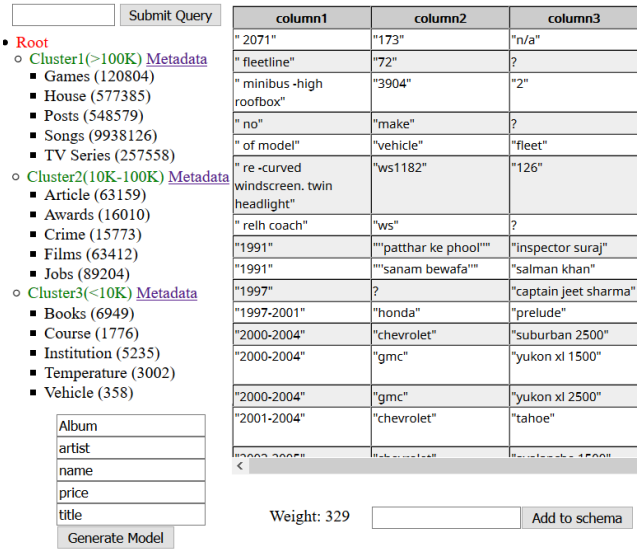

Fig. 1. User Interface. See live demo at https://youtu.be/aStjuPrVcsw

II. Related work

[Cafarella et al., 2007] investigated the problem of schema inference from a structured data corpus with incomplete or missing attribute names. Our algorithms, by contrast, summarize a large-scale structured dataset by crystallizing and ranking the main data tables, the majority of which have metadata. In another work, [Cafarella et al., 2008] create a database engine `AcsDB` (attribute correlation statistics database) that ingests Web tables. They describe several use cases, schema auto-complete (given set of attributes, suggest a table with these attributes, attribute synonym finding (given an attribute, suggest similar attributes), and join-graph discovery (a data exploration tool that represents the table corpus as a graph with nodes corresponding to tables and edges related to pairs of relations sharing a given attribute). Our work here addresses a different problem of automatic summarization of a large-scale structured dataset and constructing its *meta-schema* with major object clusters.

References

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *WWW*.

[Cafarella et al., 2008] Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., and Zhang, Y. (2008). Webtables: exploring the power of tables on the web. *VLDB*.

[Cafarella et al., 2007] Cafarella, M. J., Suciu, D., and Etzioni, O. (2007). Navigating extracted data with schema discovery. In *WebDB*. Citeseer.