SourceNote: Discovering Invaluable Data Sources in Volume

Amol Deshpande
University of Maryland
amol@cs.umd.edu

Xin Luna Dong
Google Inc.
lunadong@google.com

Lise Getoor
University of California, Santa Cruz
getoor@soe.ucsc.edu

Theodoros Rekatsinas
University of Maryland
thodrek@cs.umd.edu

Divesh Srivastava
AT&T Labs Research
divesh@research.att.com

ABSTRACT

Data is becoming a commodity and integrating data from multiple data sources has tremendous value for many public and enterprise application domains. However, the number of data sources has risen rapidly due to recent developments in data publishing and availability over the web. The proliferation of services such as cloud-based data markets has facilitated the collection, publishing and trading of data. Furthermore, the adoption of open data policies both in science and government has increased the amount of open access data without restrictions or fees promoting the idea that data should be universally available. However, data sources are typically heterogeneous in their focus and content, they often providing duplicate and conflicting information and also vary significantly in terms of the accuracy and the timeliness of the data they provide. When the number of data sources is large, humans have a limited capability of extracting accurate estimates of source authoritativeness and quality.

In this paper we explore the problems of appraising, managing and reasoning about heterogeneous data sources for data integration and collective analysis. Given the sheer number of available data sources, analysts must (1) identify sources that potentially satisfy the needs of their applications with few effective clues about the content and quality of the sources, (2) repeatedly invest many man-hours in assessing the eventual usefulness of these sources by manually investigating their content or integrating subsets of them and evaluating the actual benefit of the integration result for their application. We propose SourceNote, a vision for a data source management system that could dramatically ease the *Identify-Evaluate-Integrate* interaction loop that many analysts follows today to discover invaluable sources for their tasks.

1. INTRODUCTION

Describe how analyzing multiple data sources is part of modern applications (give examples from aggregators like datasift clients, OSI, political scientists).

Talk about redundancy of sources, other characteristics that make it hard for the user to identify useful sources (have plots from GDELT). Show content diversity.

Conclude with paper structure: (1) describe challenges, including (i) how to formally define quality, (ii) how to reason about the estimate the quality of arbitrary sets of sources, (iii) how to support diverse tasks from different users and how the user will interact with the system, (2) present overview of SourceNote and describe (i) support for heterogeneous sources diverse tasks, (ii) quality discovery module, dependency discovery and representation, (iii) query answering module. (3) related work conclusions.

2. CHALLENGES

2.1 Reasoning about the Quality of Sources

Describe quality metrics and example of each quality metric: (i) coverage, (ii) freshness, (iii) timeliness, (iv) accuracy, (v) bias.

2.2 Estimating the Integration Quality

Build quality profiles for individual sources and present how they can be combined using probabilities.

Challenge of dependent sources.

2.3 Supporting Diverse Integration Tasks

Present challenge of supporting diverse tasks. Give examples of diverse analytics tasks, structured vs. unstructured data.

2.4 User Interaction

different cost functions (i) monetary, (ii) amount of data. How can a user identify and specify the right quality for her application.

What are the right visualizations and interaction schemes.

3. FRAMEWORK OVERVIEW

Present architecture framework. Core module *knowledge graph and correspondence graph*. Modules:

1. Query engine: free-text queries mapped to entities and concepts. Quality metrics specified, cost specified, time-points of interest.

2. Correspondence graph: Go from entities to set of relevant sources (ground the domain of interest).

3. Source quality estimation: Given quality annotations of correspondence graph estimate the quality for future time points.

4. Source selection module: Find source selection solutions on pareto frontier. Given quality and cost constraints find solutions on pareto frontier.

5. Query optimizer: find diverse solutions on pareto frontier. Speculative query answering for neighborhood on pareto frontier.

6. Result visualization: Present results to user.

3.1 Supporting Diverse Integration Tasks

Analyze concepts of knowledge graph and correspondence graph. Describe construction of correspondence graph. How can you compute quality.

3.2 Estimating the Quality of Integration

Describe how a correspondence graph can be extended with factor graphs and describe compilation techniques to compute quality of integration. Learn dependencies for different metrics. The domain corresponds to unions of concepts and instances.

3.3 Query Processing

Descibe finding solutions on pareto frontier, describe how to diversify solutions
How to visualize solutions (coverage, freshness source char plot from SIGMOD 2014), how to explore the pareto frontier and neighborhood of solutions; a two level approach: **Level 1:** return ranked list with “distant" solutions on the pareto curve just mention quality characteristics, number of sources and cost. Present an average characteristic vector, no names or individual characteristics of sources, **Level 2:** once the user selects a solution from the ranked list, then present bubble graph based on characteristics of sources the two dimensions should be "concept focus and instance focus", bubble size should be source size (whatever the source is reporting). The user can click to remove a source from the solution and the characteristics (overall quality, cost) of the solution should be updated. Apart from removing sources neighboring solutions of the pareto curve should be presented in a separate list.

4. RELATED WORK

5. CONCLUSIONS