Finding Quality in Volume: The Challenge of Discovering Valuable Sources for Integration

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

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

Xin Luna Dong
Google Inc.
lunadong@google.com

Lise Getoor
UC, Santa Cruz
getoor@soe.ucsc.edu

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

ABSTRACT

Data is becoming a commodity of tremendous value for many applications, leading to a rapid increase in the number of open access data sources and services, such as cloud-based data markets and data portals, that facilitate the collection, publishing and trading of data. Data sources are typically heterogeneous in their content, the quality of data they provide and the fees they may require for accessing their entries. However, when the number of data sources is large, humans have a limited capability of reasoning about the actual quality of sources the trade-off between the benefits and costs of acquiring and integrating sources. Given the sheer number of available data sources, analysts must (1) identify sources that are relevant to their integration task, (2) discover sources that potentially satisfy the quality and budget requirements of their applications with few effective clues about the quality of the sources, (3) repeatedly invest many man-hours in assessing the eventual usefulness of data sources by manually investigating their content or integrating subsets of them and evaluating the actual benefit of the integration result for their application. In this paper we explore the problems of appraising the quality of data sources and identifying the most beneficial sources for diverse integration tasks. We introduce our vision for a new data source management system that automatically assesses the quality of data sources based on a collection of rigorous data quality metrics and enables the automated discovery of valuable sources for user specified integration tasks. We argue that the proposed system can dramatically ease the *Discover-Appraise-Evaluate* interaction loop that many analysts follow today to discover beneficial sources for their tasks.

1. INTRODUCTION

In the last few years, the number of data sources available for integration and analysis has risen because of the ease of publishing data on the Web, the proliferation of services that facilitate the collection and sharing of data (e.g., Google Fusion Tables), and the adoption of open data access policies both in science and government. This deluge of data has enabled small and medium enterprises as well as data enthusiasts to increasingly acquire and analyze data from multiple data sources.

However, the sheer number of available data sources makes it challenging for a user to identify sources that are truly beneficial to her integration or analysis task. First, many sources provide erroneous or stale data entries that can be detrimental to the quality of integration [7, 13] or provide duplicate and redundant data making hard to identify the unique information a source is providing [4, 13]. Second, sources exhibit significant heterogeneity in representation of stored data (e.g., the schema they follow), making it challenging for a user to discover all relevant sources for her integration task [5]. Finally, acquiring and integrating data comes with a monetary and computational cost, and hence, integrating every available source may not be worthwhile or even feasible due to budget constraints. These challenges give rise to the natural questions of (i) how can one specify the value of data in a rigorous fashion and (ii) how can one identify the most beneficial data sources for arbitrary integration tasks.

In recent work [8, 15] we showed that given a fixed data domain, the benefit of integration can be specified using rigorous quality metrics such as *coverage*, *accuracy* and *freshness*. We also introduced the paradigm of *source selection* to reason about the benefits and costs of acquiring and integrating data. Yet our proposed techniques focus on pre-defined integration tasks and do not provide support for a diverse set of users to specify their integration task. Moreover, our approach requires that a user knows exactly which quality metric is of importance to her or what her desired trade-off between multiple metrics is. This last requirement is rather unrealistic as users rarely know what is the right trade-off between different quality metrics.

In this paper we introduce our vision for a *source management system* that will enable users to specify heterogeneous integration tasks as keyword queries and allow them to interactively discover the most important sources for their specified task. Given an extensive collection of rigorous data quality metrics, we envision a system that automatically discovers the quality of different data sources and extends our idea of source selection to enable discovery and exploration of data sources. A key characteristic of the proposed system is to facilitate users understand which quality metrics are important for their integration task and enable them to discover sources whose integration result will maximize the desired metrics under any specified budget constraints.

The remainder of the paper is organized as follows. In Section 2 we discuss the key requirements and challenges in building a data source management system, including, defining and computing multiple quality metrics that characterize the content of sources, supporting diverse integration tasks expressed as keyword queries,

and enabling the interactive exploration of data sources. Then, Section 3 presents and overview of our proposed system, introduces the different modules of the system and proposed techniques for addressing the aforementioned challenges. Finally, Section 4 discusses related work and Section 5 concludes the paper.

2. DISCOVERING VALUABLE SOURCES

In this section we describe two tasks that require integrating data from multiple sources. Our goal is to identify common characteristics across diverse tasks and building upon these introduce the main operations that a source management system should be able to support for discovering valuable sources. Finally we describe the challenges in each of these operations.

2.1 Motivating Examples

Our goal is to demonstrate how data from multiple sources can be combined to perform diverse tasks and what are the main characteristics that one can use to fully specify diverse integration tasks. The first task we consider corresponds to combining data from multiple news papers to validate a new theory in political sciences, while the second one corresponds to a company converting social media activity into early signals to spot emerging trends for Fortune 500 companies, news organizations and financial institutions.

We consider the Global Database of Events, Languages and Tone (GDELT) [12] where news articles from thousands of news domains (sources) are aggregated in a single repository. The articles are analyzed and events, defined as pairwise interactions between actors, are extracted from them. Actors typically correspond to well-known organizations, including countries or international organizations. The extracted events are quite diverse ranging from economic agreements between companies to violent acts between parties and are associated with certain locations, actors (corresponding to organizations), and a description containing the political views of the actors and a characterization of the event. Such a repository can be used for diverse analytics tasks. We focus on a scenario where analysts integrate events mentioned in a diverse set of news media sources and analyze them collectively to detect patterns and evaluate new theories. In particular, we focus on the work by Schutte et al. [16]. In this work, the authors focus on finding causal relationships in spatiotemporal event data and use data from GDELT to build test cases for validating their theory. In particular, they are interested in examining how the civilian assistance to US forces changed in response to indiscriminate insurgent violence in Iraq. They focus on a specific time-window from 2003 to 2010, consider events in specific locations in Iraq and consider actors corresponding to specific ethnic groups in Iraq.

Next, we consider the company TrendSpottr [2] and its partnership with Datasift [1]. Datasift is a company that extracts social media data from multiple sources, including Twitter, Facebook, Blogs etc., and offers a common API for accessing dozens of different data sources in real-time. TrendSpottr analyzes real-time data streams such as Twitter and Facebook to spot emerging trends at their earliest acceleration point. They require access to historical data, demographic information and geo location information as all of this data is important to the predictive model used by the company. Typical tasks supported by TrendSporttr are viral news discovery corresponding to different locations and specific organizations, brand and reputation monitoring, ranging from company acquisitions to sports, market predictions and many others.

In both scenarios, we see that specifying an integration task involves providing information about its *location*, a set of exact instances or types of *organizations* or *people* involved in it and a *time window* associated with the task (this can be either a real-time task or a historical task over a past time-window). Moreover, it is easy to see that different types of data quality are implicitly relevant to each of the aforementioned tasks. For example, *completeness* is important for the first task but timeless is not crucial as delayed mentions of events will not affect the quality of the analysis. On the other hand, *timeliness* is very important in the second application. Analyzing sources that exhibit significant delays beats the purpose of the application which is early detection of trends.

2.2 Key Operations

Building upon the previous examples, we present the main operations that a source management system should support for discovering valuable sources. First, we consider the input to such a system. We argue that the system should ask a user to specify a desired integration task providing the following information: (i) a keyword description of the integration task corresponding to a collection of mentions to either specific objects or types (concepts) of locations, organizations and people, (ii) a selection of relevant data quality metrics from a list of supported metrics, and (iii) a desired budget characterizing the amount of money the user can afford for acquiring data. Given these specifications as input a source management system should perform the following operations:

1. **Discover.** Given the keyword description of the integration task the system should automatically determine which sources are relevant to the task by mapping the content of the task description to the content of sources.

2. **Appraise.** Once the system has discovered a set of relevant sources to the user's integration task it should detect automatically subsets of sources that if integrated together the quality of integrated data will be maximized with respect to the specified quality metrics and the imposed budget constraints. The system should identify multiple solutions that correspond to different trade-offs among the quality metrics specified by the user in a try to diversify the solutions presented to the user in the last phase. Exploring different quality metric trade-offs is necessary for the user to understand what are the exact quality characteristics required for her task.

3. **Evaluate.** The solutions discovered in the previous phase should be presented to the user together with a concise description of their quality characteristics as well as a description of the data sources included in each of them. Moreover, the user should be able to explore the neighborhood of a solution by removing sources or examining solutions with similar characteristics that contain different sources.

2.3 Challenges

Next, we discuss the main challenges in each of the operations presented above.

2.3.1 Diverse Integration Tasks

The first challenge is supporting integration tasks over heterogeneous data domains. As illustrated in the previous examples, the different concepts or entities associated with an integration task can vary significantly. Moreover, the data from the sources can be both structured and unstructured. Therefore, a data source management system should be able to reason about the semantic content of different types of data ranging from tables to free-text. Moreover, the system should be able to analyze the content of the user-provided task description without requiring a specific schema or considering only a pre-specified set of terms. Once the content is analyzed it should be semantically matched to the content of the

available sources included in the system. To fulfill these requirements one should be able to support scalable search over different types of data (e.g., structured and unstructured) while following an open-domain assumption and not restricting user queries to a pre-specified vocabulary.

2.3.2 Data Source Quality

The second challenge is defining the quality of data and being able to assess the quality of different sources in an automated way. Traditionally, the quality of data sources had been measured using the amount of erroneous information provided by the source. However, as discussed earlier there are multiple quality metrics that might be of interest to a user.

Given an integration task $\mathcal{I}$ let its data universe $\mathcal{U}(I)$ to be the set of all objects that are relevant to task $\mathcal{I}$. Notice, that $\mathcal{U}(I)$ is not fully known but only partial information is available for this universe via the data sources. We assume that each object $r \in \mathcal{U}(I)$ has a set of attribute values denoted by $r.A$. For example these attributes can correspond to a location or a type or instance of an organization as described in the previous examples. Moreover, we assume that either the objects in $\mathcal{U}(I)$ or their values may change over time. Given a set of sources S relevant to task I and an integration model F we have that $F(S) \subseteq \mathcal{U}(I)$, where $F(S)$ denotes the set of objects extracted after integrating all sources in S. Using this notation we define the following metrics:

Coverage. The coverage of S with respect to $\mathcal{U}(I)$ can be defined as the probability that an object chosen from $\mathcal{U}(I)$ uniformly at random will be present in $F(S)$. Notice that we do not require that the attributes of the object $r.A$ are correct.

Accuracy. The accuracy of S with respect to $\mathcal{U}(I)$ can be defined as the probability that an object chosen from $F(S)$ uniformly at random is correct with respect to $\mathcal{U}(I)$. The latter means that the object must be present in $\mathcal{U}(I)$ and all its attributes should have the correct values capturing errors due to both noisy and stale data.

Timeliness. The timeliness of S with respect to $\mathcal{U}(I)$ can be defined as the cumulative probability distribution of a change in the objects of $\mathcal{U}(I)$ being reflected in $F(S)$ with a delay of at most t time units. Notice that this last quality metric does not correspond to a single number but an entire distribution characterizing the delays the source exhibits.

Notice that all definitions presented above, compare the content of a source or a set of sources with the data universe corresponding to an integration task. However, the actual content of the universe is unknown and only partial information is available via content samples from the sources. Given this, the main challenge becomes combining the available source samples to extract a sufficient view of the data universe. Another challenge is that one should be able to compute the aforementioned quality metrics efficiently for any set of sources. Computing the quality of any possible set of sources in advance is obviously prohibitive. However, for certain cases [8, 15], one can estimate the overall quality for any set of sources by building offline quality profiles for each individual source and then combining those during source selection to estimate the overall quality for an arbitrary set of sources. The high-level intuition behind this approach is that each of the aforementioned quality metrics is associated with a random variable following a specific probability distribution (i.e., a Bernoulli distribution for coverage and accuracy and an empirical distribution for timeliness). If the sources are assumed to be independent the corresponding random variables are also independent, and hence, the probabilities corresponding to the quality of a set of sources can be computed efficiently using the decomposable disjunction formula. For example, the overall coverage for a set of sources S_1 and S_2 with individual coverages $C(S_1) = 0.6$ and $C(S_2) = 0.7$ corresponds to the probability that an item from $\mathcal{U}(I)$ is either covered by S_1 or covered by S_2 and is $C(S_1, S_2) = 1 - (1 - 0.6)(1 - 0.7) = 0.88$.

Figure 1: Framework overview.

In reality, however, sources are far from independent [3, 6], as they exhibit overlaps, copying relationships and/or may contradict each other. These relationships make the quality random variables for each source dependent. Estimating the aforementioned quality metrics under the presence of dependencies imposes a major challenge as: (i) it requires extracting the source dependencies from available source samples and (ii) devising efficient techniques for computing the probability of the overall quality random variables during query evaluation.

2.3.3 Interactive Evaluation

As mentioned above, users should be able to specify a certain budget for their integration task. This can be either a monetary budget, limiting the amount of data that can be acquired, or a budget on the number of sources that a source selection solution should contain. The latter is particularly useful when a user wants to verify the content of sources manually. While specifying a constraint on the integration budget is natural to users, specifying a constraint on the data quality with respect to the aforementioned metrics is rather intricate due to the interdependencies between the different metrics and we expect that not even expert users know the exact data quality requirements of a task. Consider for example a political scientist analyzing news papers articles to forecast interesting events. Imposing a constraint on using only data sources that have exhibited zero delay with respect to certain events may reduce the accuracy of the integration result. Moreover, users may not know what is the feasible level of quality given their budget and what are the trade-off between solutions that focus on maximizing the different quality metrics in isolation. For example, selecting sources optimizing for accuracy may lead to an integration result of low coverage.

So a major challenge for a source management system is to guide the user through the different source selection solutions that satisfy the user's budget and help her understand the trade-off between the integration quality achieved by different solutions. We argue that this is feasible only through an interactive process where the user will be able to explore the feasible solution space following suggestions of the source management system. This will enable users to understand the interdependencies between the different quality metrics with respect to their integration task and identify the particular solution that suits their application. The latter raises the following challenges: (i) how can a user explore the solution space efficiently, (ii) how can a source management system present the quality profile for a set of sources in a concise and meaningful way, and (iii) what are the right hints that the system should present to the user to facilitate the exploration of the solution space.

3. SYSTEM OVERVIEW

In this section we propose a preliminary design that aims to address the challenges specified in the previous section. An overview of the proposed system is show in Figure 1.

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 domains of interest as a clusters of concepts and instances).

3. Source quality estimation: Given quality annotations of correspondence graph estimate the quality for future time points. The annotations will be either on edges from domain clusters to sources (for independent sources) or factor graphs at the cluster level for dependent sources.

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 underlying quality characteristics of each source (i.e., compare content of sources).

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. For each domain cluster: the domain corresponds to a union of concepts and instances.

3.3 Query Processing

Descibe finding solutions on pareto frontier, describe how to diversify solutions. See papers on probing pareto frontier.
How to visualize solutions (coverage, freshness source bubble 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 correspond to 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 initial solution on the pareto frontier should be presented in a separate list.

4. RELATED WORK

Integrating data from multiple sources is essential in a growing number of application domains, including large scale enterprises that own many data sources [9], collective intelligence, which aggregates the shared information from a diverse set of sources [10, 14], and targeted data analytics where data of Web and social media are heavily used [11]. Analyzing multiple data sources collectively can significantly enhance the value of data. For example, with more sources, the completeness of the integrated data can be increased; in the presence of inconsistencies, the correctness of the integrated data can be improved by leveraging the collective wisdom. Such quality improvements allow for more advanced data analysis and can bring a big *gain*. Focus on integrating all sources and focus on accuracy alone.

5. CONCLUSIONS

6. REFERENCES

[1] Datasift. `http://datasift.com`.

[2] Trendsporttr. `http://trendspottr.com`.

[3] L. Berti-Equille, A. D. Sarma, X. Dong, A. Marian, and D. Srivastava. Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In *CIDR*, 2009.

[4] M. Bronzi, V. Crescenzi, P. Merialdo, and P. Papotti. Extraction and integration of partially overlapping web sources. *PVLDB*, 6(10):805–816, 2013.

[5] A. Das Sarma, L. Fang, N. Gupta, A. Halevy, H. Lee, F. Wu, R. Xin, and C. Yu. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’12, pages 817–828. ACM, 2012.

[6] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.*, 2(1):550–561, Aug. 2009.

[7] X. L. Dong, A. Halevy, and C. Yu. Data integration with uncertainty. *The VLDB Journal*, 18(2):469–500, Apr. 2009.

[8] X. L. Dong, B. Saha, and D. Srivastava. Less is more: selecting sources wisely for integration. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB’13, pages 37–48. VLDB Endowment, 2013.

[9] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB ’06, pages 9–16. VLDB Endowment, 2006.

[10] T. Hua, C.-T. Lu, N. Ramakrishnan, F. Chen, J. Arredondo, D. Mares, and K. Summers. Analyzing civil unrest through social media. *Computer*, 46(12):80–84, 2013.

[11] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan. Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, pages 1470–1473. ACM, 2013.

[12] K. Leetaru and P. Schrodt. Gdelt: Global data on events, language, and tone, 1979-2012. *Inter. Studies Association Annual Conf.*, 2013.

[13] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: is the problem solved? In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB’13, pages 97–108. VLDB Endowment, 2013.

[14] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1041–1052. International World Wide Web Conferences Steering Committee, 2013.

[15] T. Rekatsinas, X. L. Dong, and D. Srivastava. Characterizing and selecting fresh data sources. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 919–930. ACM, 2014.

[16] S. Schutte and K. Donnay. Matched wake analysis: finding causal relationships in spatiotemporal event data. *Political Geography*, 41:1–10, 2014.