Poster Paper

Context-Dependent Quality-Aware Source Selection for Live Queries on Linked Data

Barbara Catania
DIBRIS
University of Genova
Genova,Italy
barbara.catania@unige.it

Giovanna Guerrini
DIBRIS
University of Genova
Genova,Italy
giovanna.guerrini@unige.it

Beyza Yaman
DIBRIS
University of Genova
Genova,Italy
beyza.yaman@dibris.unige.it

ABSTRACT

Source selection deserves attention for live query processing over distributed, poorly controlled data sources since it is the key to produce the best available information, in terms of relevance, trustness, and freshness, as query result. In this paper, we present an approach taking into account context-dependent data quality, according to different dimensions, during source selection, with the aim of selecting not only the most relevant but also the highest quality sources.

1. INTRODUCTION & BACKGROUND

In the last decade significant amount of Linked Data has been published to construct a global data space. However, it is still difficult to benefit from published data in an effective way because identifying the sources containing the most valuable results for a query is a non-trivial task. We propose an approach to select sources taking into account not only relevance but also quality in a context-dependent way.

Our approach fits in the general vision of Rekatsinas *et al.* [2] for a data source management system that enables users to discover the most valuable data sources for their applications. To characterize data source value, different *quality indicators* [1] can be used to assess the different quality dimensions (accuracy, freshness, completeness, etc), relying on data content, metadata (such as update times) and explicit user feedbacks. Data quality, however, may be different if assessed with reference to a geographical area, historical period, or type of content. As a result, it is not only impossible to assess quality in an absolute way, but it is difficult as well to assess a single quality dimension independently from the *context*. Quality indicators would then be associated with data according to the different contexts.

We adopt the notion of context proposed in [2] in terms of context clusters. A context cluster describes the data domain corresponding to a data collection and it is specified as a conjunction of a set of concept classes and a set of instances. In our approach, context clusters can specify information about *what* (the type of the described information), *where* (spatial location), *when* (temporal location), and *why* (data motivations).

Each data item (i.e., each RDF triple) is associated with (one or more) context clusters. A source can include data items from different contexts. A (data source, context) pair intensionally characterize a data collection, consisting of the set of triples in the source associated with the context. Quality indicators are associated, for the relevant quality dimension, with such data collections. A data source can thus exhibit different quality degrees, resulting in different indicator values, according to the different contexts.

Given a user query, the approach consists of four tasks: (i) context-dependent quality aware source selection to devise the most relevant and highest quality sources according to the query and its context; (ii) feedback from the user on the results obtained by the query evaluation; (iii) refinement/update of the data quality indicators according to such feedback; (iv) update of the auxiliary structures employed for source selection according to the refinement in (iii).

The proposed approach relies on two main notions, that are combined in an original way:

1. *Named Graphs.* Named Graphs [1] are useful structures for hierarchically composing subgraphs and building nested graphs. They allow to represent and exchange metadata.

2. *Data Summaries.* Data summaries [3] have been proposed to efficiently determine which sources may contribute answers to a query in live distributed query systems. They approximately describe the data provided by a source in an aggregated form, in much more detail than schema-level indexes. The QTree, specifically, is a data summary over Linked Data sources, seeing the data items (RDF triples) in the sources as points of a three dimensional numerical space, by applying hash functions to the triple components. Like the R-tree, a QTree is a tree structure consisting of nodes defined by minimal bounding boxes (MBBs). An MBB describes the multidimensional region in the data space that is represented by the node the subtree underneath. Leaf nodes in a QTree, however, rather than containing the data items that are contained in their MBBs as in R-trees, are buckets containing statistical information (e.g., count) that approximate the data items contained in their MBBs.

2. SOURCE SELECTION EXPLOITING CONTEXT-DEPENDENT QUALITY

Figure 1 provides a graphical overview of the approach. We first briefly describe the steps in the source selection process and then discuss its main components. First, the

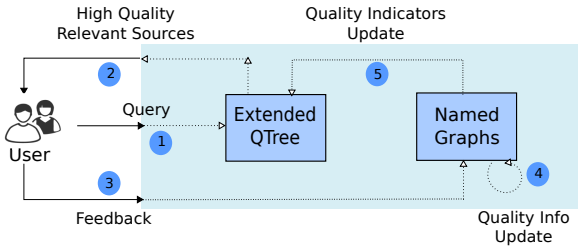

Figure 1: Overview of the Approach

user query is submitted to the system (Edge 1) and a look-up is performed on an extended QTree, returning to the user a ranked list of high quality relevant sources (Edge 2). The extended QTree allows the retrieval of potentially relevant sources with quality indicators, for given quality dimensions in a given context, above a given threshold. Once query results are returned to the user, her feedback (if any, Edge 3) on the obtained results is considered, resulting in an update and refinement of the quality metadata associated with the sources according to the context (Edge 4). Such context-dependent metadata are maintained making use of Named Graphs. Named Graphs mantain detailed quality indicators and metadata, so that this knowledge is shared and easily accessible. Such detail quality information is the basis on which the numeric indicators in the extended QTree are computed. Finally, context-dependent numerical quality indicators are incrementally updated in the extended QTree according to the new metadata (Edge 5).

Query. The query is a conjunctive SPARQL query associated with context information and quality thresholds. Conjunctive SPARQL queries consist of so-called basic graph patterns (BGPs), i.e., sets of triple patterns in the (*subject*, *predicate*, *object*) form possibly containing variables. We assume that the BGPs come with context information in the form of context clusters. Moreover, thresholds can be specified with respect to one or several among the supported quality dimensions. Only sources associated with indicators with values above the thresholds for the specified dimensions will be considered as result, while the other ones are discarded.

Named Graph. We exploit Named Graphs to associate quality related metadata and indicators with data sources, in a context-dependent way, in order to make this knowledge shared and easily accessible. Context clusters are organized in a hierarchical structure enabling the support of different detail levels. For instance, referring to spatial location context clusters, the {`Rome`} cluster is more specific than the {`Central Italy`} cluster. A data source may contain data items from various contexts. For instance, a data source containing information about touristic attractions in central Italy may contain data items associated with the {`Museum`} (*what*) cluster as well as items associated with the {`Rome`} (*where*) cluster as well as items associated with the {`Florence`} (*where*) cluster.

A data source can be associated in the Named Graphs with different quality profiles corresponding to the different contexts of its triples. Inside Named Graphs, indeed, quality-related metadata w.r.t. different quality dimensions are attached to each data collections characterized by data source and context. This allows a context-dependent quality assessment. Referring to the above mentioned data source, its freshness may be different according to the {`Rome`} or to the {`Florence`} context, and, similarly its trustness may be different according to the {`Museum`} or {`Rome`} context. We rely on metadata standard vocabularies to describe quality-related metadata [1]. For instance, freshness can be described by using access (`dc:date`) and creation dates (`dc:created`), while trustness can be related to the creator of the source (`dc:publisher`).

Extended QTree. The QTree index is extended to consider context as a fourth component associated with triples and to associate context-dependent numerical quality indicators with data collections. Contextualized data items are now seen as point of a four dimensional numerical space, since context clusters are hashed to a numerical value as the other triple components. Each four dimensional MBB, moreover, is now associated with a quality range for each of the quality dimension. Indicators are separately computed for each quality dimension from the quality information (metadata and indicators) in the Named Graphs. The approach is flexible, in that the employed indicators may be different from source to source.

Indicator ranges are used during source selection to prune the data sources with quality below the requested thresholds. Specifically, given a four dimensional region and the thresholds for quality dimensions resulting from a query, the extended QTree is employed by descending in a child node if the region is contained in the MBB and the quality ranges intersect the corresponding thresholds. Indicators are finally also employed to rank the retrieved data sources according to their quality w.r.t. the query context.

3. CONCLUSION

In this paper we propose an approach to select relevant sources from an arbitrary, unrestricted set of distributed, poorly controlled Linked Data sources, so that queries can be processed on these sources taking into account not only their relevance to the query but also their quality, in terms of a number of dimensions, with respect to the query context.

Technically, the proposed approach relies on the use of nested Named Graphs to associate quality metadata with data source according to different contexts and at different granularity levels, and on extended QTree enabling efficient source selection, not only relying on relevance, but also on context-based quality indicators.

4. REFERENCES

[1] C. Bizer. *Quality Driven Information Filtering: In the Context of Web Based Information Systems*. VDM Publishing, 2007.

[2] T. Rekatsinas, X. L. Dong, L. Getoor, and D. Srivastava. Finding quality in quantity: The challenge of discovering valuable sources for integration. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2015.

[3] J. Umbrich, K. Hose, M. Karnstedt, A. Harth, and A. Polleres. Comparing Data Summaries for Processing Live Queries over Linked Data. *World Wide Web*, 14(5-6):495–544, 2011.