Web Semantics: Science, Services and Agents on the World Wide Web 26 (2014) 1–28

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Querying a messy web of data with AVALANCHE

Cosmin Başca *, Abraham Bernstein

Dynamic and Distributed Information Systems, Department of Informatics, University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 11 February 2011
Received in revised form
6 February 2014
Accepted 8 April 2014
Available online 24 April 2014

Keywords:
Federated SPARQL
RDF distribution messiness
Query planning
Adaptive querying
Changing network conditions

ABSTRACT

Recent efforts have enabled applications to query the entire Semantic Web. Such approaches are either based on a centralised store or link traversal and URI dereferencing as often used in the case of Linked Open Data. These approaches make additional assumptions about the structure and/or location of data on the Web and are likely to limit the diversity of resulting usages.

In this article we propose a technique called AVALANCHE, designed for querying the Semantic Web without making any prior assumptions about the data location or distribution, schema-alignment, pertinent statistics, data evolution, and accessibility of servers. Specifically, AVALANCHE finds up-to-date answers to queries over SPARQL endpoints. It first gets on-line statistical information about potential data sources and their data distribution. Then, it plans and executes the query in a concurrent and distributed manner trying to quickly provide first answers.

We empirically evaluate AVALANCHE using the realistic FedBench data-set over 26 servers and investigate its behaviour for varying degrees of instance-level distribution "messiness" using the LUBM synthetic data-set spread over 100 servers. Results show that AVALANCHE is robust and stable in spite of varying network latency finding first results for 80% of the queries in under 1 s. It also exhibits stability for some classes of queries when instance-level distribution messiness increases. We also illustrate, how AVALANCHE addresses the other sources of messiness (pertinent data statistics, data evolution and data presence) by design and show its robustness by removing endpoints during query execution.

1. Introduction

With the advent of the Semantic Web, a Web-of-Data is emerging interlinking ever more machine readable data fragments represented as RDF documents or queryable semantic endpoints. It is in this ecosystem that unexplored avenues for application development are emerging. While some application designs include a Semantic Web data crawler, others rely on services that facilitate access to the Web-of-Data either through the SPARQL protocol or various APIs like the ones exposed by *Sindice*[1] or *Swoogle*.[2] As the mass of data continues to grow – Linked Open Data [5] accounts for 27 billion triples as of January 2011[3] – the scalability factor combined with the Web's uncontrollable nature and its heterogeneity will give rise to a new set of challenges. A question marginally addressed today is how to support the same messiness in querying the Web-of-Data that gave rise to the virtually endless possibilities of using the traditional Web. In other words: *How can we support querying the messy web of data whilst adhering to a minimal, least-constraining set of principles that mimic the ones of the original web and will – hopefully – support the same type of creative flurry?*

Translating the guiding principles of the Web to the Web-of-Data proposes that we should use a single communications protocol (i.e. HTTP with encoded SPARQL queries) and use a common data representation format (some encoding of RDF), which allows embedding links. In addition, it implicitly proposes that:

(a) we cannot assume any (or control the) distribution of data to servers,
(b) there is no guarantee of a working network,
(c) there is no centralised resource discovery system (even though crawled indices akin to Google in the traditional web may be provided),
(d) the size of RDF data no longer allows us to consider single-machine systems feasible,
(e) data will change without any prior announcement,
(f) there is absolutely no guarantee of RDF-resources adhering to any kind of predefined schema, being correct, or referring/link-

* Corresponding author. Tel.: +41 44 635 4318.
E-mail addresses: basca@ifi.uzh.ch, cosmin.basca@gmail.com (C. Başca), bernstein@ifi.uzh.ch (A. Bernstein).

1 http://swoogle.umbc.edu/.
2 http://sindice.com/.
3 http://www4.wiwiss.fu-berlin.de/lodcloud/state/#domains.

http://dx.doi.org/10.1016/j.websem.2014.04.002
1570-8268/

ing to other existing data items—in other words: the Web-of-Data will be a mess and "this is a feature not a bug".

As an example, consider the life sciences domain: here information about drugs, chemical compounds, proteins and other related aspects is published continuously. Some research institutions expose part or all of their data freely as RDF dumps relying on others to index it as in the cases of the CheBi[4] and KEGG[5] datasets, while others host their own endpoints like in the case of the Uniprot dataset.[6] Hence, anybody querying the data will have:

- no control over its distribution, i.e. different copyright and intellectual property policies may prevent access to downloading part or the entire dataset but permit access to it on a per-query basis with potential restrictions like time and/or quota limits,
- no guarantees about the availability and network connectivity of the information sources, i.e. some institutions move repositories or change access policies, resulting in server unavailability,
- no guarantees about content stability as data changes continuously due to scientific breakthroughs/discoveries, and a plethora of schemas are used, i.e. some sub-disciplines may favour dissimilar but overlapping attributes describing their results, have differing habits about using same-named attributes, and use a diversity of taxonomies with varying semantics.

Often-times problem domains and researchers' questions span across several datasets or disciplines that may or may not overlap. Even in the light of this messiness, the data about drugs, chemical compounds, proteins, and their interrelations is queried constantly resulting in a strong need to provide integrated and up-to-date (or current) information.

Several approaches that tackle the problem of querying the entire Web-of-Data have emerged lately, and most adhere to the explicit principles. They do, however, not address the implicit principles. One solution, *uberblic.org*,[7] provides a centralised queryable endpoint for the Semantic Web that caches all data. This approach allows searching for and joining potentially distributed data sources. It does, however, incur the significant problem of ensuring an up-to-date cache and might face crucial scalability hurdles in the future, as the Semantic Web continues to grow. Additionally, it violates a number of the implicit principles locking-in data. Furthermore, as Van Alstyne et al. [40] argue, incentive misalignments would lead to data quality problems and, hence, inefficiencies when considering the Web-of-Data as "one big database".

Other approaches base themselves on the guiding principles of Linked Open Data publishing and traverse the LOD cloud in search of the answer. Obviously, such a method produces up-to-date results and can detect data locations only from the URIs of bound entities in the query. Relying on URI structure, however, may cause significant scalability issues when retrieving distributed datasets, since (i) the servers dereferenced in the URI may become overloaded and (ii) limits the possibilities of rearranging (or moving) the data around by binding the id (*i.e.*, URI) to its storage location. Just consider for example the *slashdot effect*[8] on the traditional web. Finally, traditional database federation techniques have been applied to query the Web-of-Data. One of the main drawbacks with traditional federated approaches stemming from their ex-ante (*i.e.*, before the query execution) reliance on *fine-grained* statistical and schema information meant to enable the mediator to build efficient query execution plans. Whilst these approaches do not assume central control over data, they do assume ex-ante knowledge about it facing robustness hurdles against network failure and changes in the underlying schema and statistics (invalidating implicit principles b and f).

In this paper, we propose Avalanche , a novel approach for querying the messy Web-of-Data which (1) *makes no assumptions about data distribution, schema, availability, or partitioning* and is skew resistant for some classes of queries, (2) provides *up-to-date results* from distributed indexed endpoints, (3) is *adaptive* during execution adjusting dynamically to external network changes, (4) *does not require detailed fine-grained ex-ante statistics* with the query engine, and (5) is *flexible* as it makes limited assumptions about the structure of participating triple stores. It does, however, assume that the query will be distributed over triple-stores and not "mere" web-pages publishing RDF. The system, as presented in the following sections, is based on a first prototype described in [3] and brings a number of new extensions and improvements to our previous model.

Consequently, Avalanche proposes a novel technique for executing queries over Web-of-Data SPARQL endpoints. The traditional *optimise then execute* paradigm – highly problematic in the Web of Data context in its original conceptualisation – is extended into an exhaustive, concurrent, and dynamically-adaptive meta-optimisation process where fine-grained statistics are requested in a first phase of the query execution. In a second phase continuous query planning is interleaved with the concurrent execution of these plans until sufficient results are found or some other stopping criteria is met. Hence, the main contributions of our approach are:

- a querying approach over the indexed Web-of-Data, without fine-grained prior knowledge about its distribution
- a novel combination of interleaving cost-based planning (with a simple cost-model) with concurrent query plan execution that delivers first results quickly in a setting where join cardinalities are unknown due to lacking ex-ante knowledge
- a reference implementation of the Avalanche system.

However, despite Avalanche's flexible and robust query execution paradigm, the method also comes with a set of limitations discussed in detail in Section 3. The main limitations are as follows:

- Avalanche does not benefit from the potential speedup exhibited by intra-plan parallelism since its current computation model does not support UNION-views,
- Avalanche can be resource wasteful for some classes of query workloads,
- embracing the WWW's uncertainties (see principles a–f), Avalanche neither guarantees result-set completeness nor the same result-set for repeated same-query executions.

Hence, Avalanche supports messiness stemming from the lack of ex-ante knowledge at various levels: data-distribution, schema-alignment, prior registration with respect to statistics, constantly evolving data, and unreliable accessibility of servers (either through network or host failure, HTTP 404's, or changes in policy of the publishers).

In the remainder we first review the relevant related work of the current state-of-the-art. The computational model is described in Section 3 while Section 4 provides a detailed description of Avalanche. In Section 5 we evaluate several planning strategies

4 http://www.ebi.ac.uk/chebi/.
5 http://www.genome.jp/kegg/.
6 http://beta.sparql.uniprot.org/.
7 http://platform.uberblic.org/.
8 http://en.wikipedia.org/wiki/Slashdot_effect.