Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

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

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

Review article

Streaming the Web: Reasoning over dynamic data

Alessandro Margara *, Jacopo Urbani, Frank van Harmelen, Henri Bal

Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

A R T I C L E I N F O

Article history:
Received 11 January 2013
Received in revised form
3 February 2014
Accepted 11 February 2014
Available online xxxx

Keywords:
Semantic Web
Stream reasoning
Survey
Stream processing
Complex Event Processing

A B S T R A C T

In the last few years a new research area, called *stream reasoning*, emerged to bridge the gap between reasoning and stream processing. While current reasoning approaches are designed to work on mainly static data, the Web is, on the other hand, extremely dynamic: information is frequently changed and updated, and new data is continuously generated from a huge number of sources, often at high rate. In other words, fresh information is constantly made available in the form of *streams* of new data and updates.

Despite some promising investigations in the area, stream reasoning is still in its infancy, both from the perspective of models and theories development, and from the perspective of systems and tools design and implementation.

The aim of this paper is threefold: (i) we identify the requirements coming from different application scenarios, and we isolate the problems they pose; (ii) we survey existing approaches and proposals in the area of stream reasoning, highlighting their strengths and limitations; (iii) we draw a research agenda to guide the future research and development of stream reasoning. In doing so, we also analyze related research fields to extract algorithms, models, techniques, and solutions that could be useful in the area of stream reasoning.

Contents

* Corresponding author.
E-mail addresses: a.margara@vu.nl, Alessandro.margara@gmail.com (A. Margara), jacopo@cs.vu.nl (J. Urbani), frank.van.harmelen@cs.vu.nl (F. van Harmelen), bal@cs.vu.nl (H. Bal).

http://dx.doi.org/10.1016/j.websem.2014.02.001

1. Introduction

The Web is highly dynamic: new information is constantly added, and existing information is continuously changed or removed. Large volumes of data are produced and made available on the Web by on-line newspapers, blogs, social networks, etc., not to mention data coming from sensors for environmental monitoring, weather forecast, traffic management, and domain specific information, like stock prices. It has been estimated that every minute on the Internet 600 videos are uploaded on YouTube, 168 millions e-mails are sent, 510,000 comments are posted on Facebook and 98,000 tweets are delivered in Twitter.[1]

In these scenarios information changes at a very high rate, so that we can identify a *stream* of data on which we are called to operate with high efficiency. In the last few years, several researchers and practitioners have proposed solutions for processing streams of information *on-the-fly*, according to some pre-deployed processing rules or queries [1]. This led to the development of various Data Stream Management Systems (DSMSs) [2] and Complex Event Processing (CEP) systems [3,4] that effectively deal with the transient nature of data streams, providing low delay processing even in the presence of large volumes of input data generated at a high rate.

All these systems are based on data models, like for example the well known relational model, which allow only a predefined set of operations on streams with a fixed structure. This allows the implementation of ad-hoc optimizations to improve the processing.

However, the Web provides streams of data that are extremely heterogeneous, both at a structural and at a semantical level. For example, a Twitter stream is radically different from a stream delivered from a news channel, not only because they are stored using different formats, but also because they contain different types of information.

Furthermore, the ability of operating on-the-fly on several of these streams simultaneously would allow the implementation of real-time services that can select, integrate, aggregate, and process data as it becomes available, for example to provide updated answers to complex queries or to detect situations of interests, to automatically update the information provided by a web site or application.

The Semantic Web is an extension of the current World Wide Web, where the semantics of information is encoded in a set of RDF statements. Currently, we are witnessing an explosion of the availability of RDF data on the Web since both public and private organizations have chosen this format to release their public data.[2]

The choice of RDF as data model, in combination with ontological languages (e.g., OWL [5]), enables the implementation of algorithms that can "reason" on existing data to infer new knowledge. Current solutions and technologies for reasoning on RDF data are designed to work on scenarios where changes occur at low volumes and frequencies, and this clashes with the dynamic nature of the streams on the Web.

To bridge this gap, a number of recent works propose to unify reasoning and stream processing, giving birth to the research field of *stream reasoning*. In 2009, stream reasoning was defined as an "unexplored yet high impact research area" [6]. After a few years of research, despite some interesting preliminary investigations in the field, we observe that the stream reasoning research area remains vastly unexplored, both from a theoretical point of view and from the perspective of systems and tools supporting it.

Contributions. In this paper, we first report some example application areas that can benefit from stream reasoning and analyze the requirements they pose. Then, we survey existing approaches in this field, and show why none of them can currently represent a complete answer to all the requirements of various application fields. Starting from this analysis, we isolate some key challenges that need to be addressed to offer full fledged tools for stream reasoning.

Finally, we elaborate a number of possible solutions to overcome the limitations of current approaches. We analyze related research fields to explore whether some topics or solutions, but also algorithms, techniques, and best practices could apply to solve open issues in stream reasoning. In doing so, we intend to illustrate the current state of the stream reasoning research area and to summarize a possible research agenda to further advance in this field.

Outline. The remainder of this paper is organized as follows. Section 2 introduces some example application scenarios for stream reasoning and analyzes the requirements they pose. Section 3 reports a high level introduction to the problem of stream reasoning by describing the research fields that are related to stream processing and reasoning. Next, Section 4 presents a survey of current proposals for stream reasoning and highlights their advantages and limitations. Section 5 extracts the open issues in our current context and Section 6 presents some possible concrete solutions to overcome these issues. Finally, Section 7 provides some conclusive remarks.

2. Motivations for stream reasoning

This section presents some motivations for the need of stream reasoning technologies. It is divided in three parts. In the first part, we present different application scenarios. In the second part, we extract some general requirements that could help identifying the main features expected from stream reasoning. In the third part, we briefly analyze these requirements, with particular focus on their mutual dependencies. Some of the scenarios listed below have already been introduced in previous works on stream reasoning [6,7]. Others are relatively new: for them, the benefits of stream reasoning technologies are discussed for the first time in this paper.

2.1. Motivating scenarios

Semantic Sensor Web. The Semantic Sensor Web (SSW) approach aims at increasing and integrating the communication between

1 http://www.go-gulf.com/blog/60-seconds.
2 http://linkeddata.org.

Please cite this article in press as: A. Margara, et al., Streaming the Web: Reasoning over dynamic data, Web Semantics: Science, Services and Agents on the World Wide Web (2014), http://dx.doi.org/10.1016/j.websem.2014.02.001