Semantic Complex Event Processing for Decision Support

Robin Keskisärkkä

Linköping University, Linköping, Sweden,
robin.keskisarkka@liu.se

Abstract. An increasing amount of information is being made available as online streams, and streams are expected to grow in importance in a variety of domains in the coming years (e.g., natural disaster response, surveillance, monitoring of criminal activity, and military planning [7, 22]). Semantic Web (SW) technologies have the potential to combine heterogeneous data sources, leveraging Linked Data principles, but traditional SW methods assume that data is more or less static, which is not the case for streams. The SW community has attempted to bring streams to a semantic level, i.e., Linked Stream Data, and a number of RDF stream processing engines have been produced [1, 4, 13, 20]. This thesis work aims at developing and evaluating techniques for creating aggregated and layered abstractions of events. These abstractions can be used by decision makers to create better situation awareness, assisting in identifying decision opportunities, structuring and summarizing decision problems, and decreasing cognitive workload.

Keywords: RDF stream processing, Semantic CEP, decision support, situation awareness

1 Introduction/Motivation

Traditionally, authorities and citizens have relied on official communication channels to achieve situation awareness, but today online communication channels are increasingly being used [11, 25]. As an example, analysis of social media has been used to assess influenza outbreaks [12], and to assist disaster relief [23]. But while humans can handle complex recognition and analysis tasks with great accuracy, performance suffers greatly as workload increases, which introduces a bottleneck for large scale data analysis tasks.

The information available as online streams is expected to grow in importance across many domains, e.g., natural disaster response, surveillance, monitoring of criminal activity, and military planning [7, 22]. Online streams are already used in domains such as electronic trading and market feed processing [18].

The Semantic Web (SW) community has attempted to bring streams to a semantic level, i.e., Linked Stream Data, but in handling continuously delivered data traditional SW technologies fall short [14]. The streaming data is characterized by being received continuously in possibly unbounded streams [4]. Since no

final answer can typically be returned, queries must be executed repeatedly as new data becomes available. The traditional methods for querying Linked Data build on the assumption of data as more or less static, but streaming data is often outdated quickly and needs to be consumed on the fly [24]. Also, the rate and quantity of the incoming data in itself may require data to be processed continuously. To tackle this problem a number of streaming engines and query language extensions, targeted at querying streaming and static RDF data, have been developed. These systems have largely been inspired by data stream management systems, e.g., CQL [2], which use query operators to isolate portions of streams based on timestamps.

The purpose of Complex Event Processing (CEP) is to (semi-)automatically create "actionable" abstractions from streams of events [16]. In order to support flexible abstractions and pattern matching of heterogeneous data streams semantics must be introduced, and Semantic CEP is targeted specifically at *semantically* interpreting and analyzing data using the Linked Data principles and SW technologies.

The abstraction task can be described as process in which sets of *low-level* events are aggregated into *high-level* events, by means of some event pattern. An event is defined here as "anything that happens, or is contemplated as happening", while a complex event is defined as "an event that summarizes, represents, or denotes a set of other events" [16].

A simple view of decision making is as a process in which an answer is selected among a set of alternatives, but a more sophisticated view includes many other aspects, such as identifying that there is a decision to be made in the first place. Empirical studies have demonstrated that human judgment and decision making relies heavily on intuitive strategies, rather than on strict reasoning rules [6]. These strategies reduce cognitive load, but also make decisions more sensitive to biases. The purpose of Decision Support Systems (DSS) is to assist decision makers in the activity of making a decision by providing a set of tools to aid decision makers in the process of modeling and analyzing data, identifying decision opportunities, imposing structure on data, and supporting the choice processes [17, Chapter 1]. Another way of supporting the decision making process is to abstract portions of a decision-making situation and to model the available information in ways which retain only the essential relationships to reduce cognitive load and problem complexity. The abstractions additionally help make knowledge easier to transfer across problems and domains.

2 Related research

The SW community is relatively new in the field of stream processing, but a considerable amount of research has previously been done on continuous query processing over streams, e.g., in projects such as the STREAM prototype developed at Stanford [2], StreamBase[1], and ESPER[2]. This research has largely

[1] http://www.streambase.com/
[2] http://esper.codehaus.org/

made up the foundation for the state-of-the-art RDF stream processing (RSP) systems that have been implemented to date.

There are no official standards for how RDF data streams are to be represented or published. In [4] an RDF stream is defined as an ordered sequence of tuples, consisting of an RDF triple and a timestamp. Triples with the same timestamp are treated as if they occurred at the same time. However, RDF streams can be represented in a number of ways, e.g., data elements can be RDF graphs rather than triples, intervals can be used instead of timestamps, and timestamps/intervals could be represented in the stream itself rather than as parts of streamed tuples [9].

In CEP streams consist of simple (atomic) and complex (composite) events. To represent events in a general sense a number of vocabularies have been proposed [21]. A common model for complex events can simplify querying and abstraction into higher level events, but it may also result in additional overhead.

2.1 State of the Art

DSSs normally have narrow application scopes and are typically not generalizable to different decision-making contexts. Over the last decade SW technologies have been used in DSS to solve tasks such as information integration and sharing, web service annotation and discovery, and knowledge representation and reasoning. For a more in-depth review of SW technologies in DSS refer to [5].

The next sections briefly describe four RDF Stream Processing (RSP) systems that represent current state of the art. These are followed by discussion of frameworks using a CEP engine in combination with Linked Stream Data.

CQELS was created to address the problem of scalable query processing of Linked Stream Data [13]. The query language is an extension of SPARQL 1.1, supporting query patterns for the defining logical and physical windows over streams. The engine is a native implementation and returns results in near real-time, and for some queries the CQELS engine outperforms similar approaches (C-SPARQL and ETALIS) by orders of magnitude [13].

C-SPARQL is an RSP engine with a query language that extends SPARQL 1.0 to support the definition of physical and logical windows over streams, as well as aggregation operators. Queries are executed periodically at a rate decided by the system, and the engine can report duplicates of results if the windows of two query executions overlap.

EP-SPARQL is a language for event processing developed for ETALIS to handle SW applications [1]. The language is based on SPARQL 1.0, but extends the language with a number of binary temporal operators. Instead of defining windows over streams functions exist to access duration, start time, and end time, and using these windows can be defined inside filter expressions.

INSTANS is a query engine capable of continuous execution of selected parts of SPARQL 1.1 Query and Update [20]. It avoids repeated computation of the same data, and makes results available immediately when a query patterns is matched. INSTANS additionally supports the detection of missing events by employing a timer that can be registered for events. Modeling of time and windows must be expressed manually in RDF (e.g. as suggested in [8]).

2.2 Framework approaches

The Streaming Linked Data (SLD) framework [3] was designed to allow publishers to stream data to a central server, where streams can be queried, stored and replayed, decorated with additional information, and republished as new streams. It assumes that streams are published using the HTTP protocol and uses a built-in RSP engine (C-SPARQL).

Another framework project is the Super Stream Collider (SSC) [19]. It to supports the registering of streams and queries, but relies instead on the CQELS engine. SSC supports streams published using Web Sockets. Web Sockets are highly efficient and allow for real-time and high speed full-duplex communication, and guarantees that data will arrive in the order it was originally streamed.

A different approach was proposed in the framework in [15]. This framework focuses on bridging the gap between Linked Data and rule-based CEP engines. This is accomplished by translating events into the format required by a specified engine, specifically for Drools Fusion[3]. The obvious benefit of this approach is that a mature and well-used CEP engine can be employed, but it also means that translations need to be provided back and forth between internal engine formats and Linked Data.

3 Problem Statement and Contributions

Current decision support systems do not assist users in making any high-level abstractions of situations, rather they only summarize or visualize data. The main hypothesis of this thesis work is that the combination of Linked Data, SW technologies, and CEP make it possible to create useful abstractions based on event patterns in real-time.

Existing vocabularies for expressing complex events will be evaluated and possibly extended. Additionally, ways of querying event streams, and representing boundaries of complex events, will be investigated. The work aims to develop ways of expressing more declarative descriptions of queries and complex events using Ontology Design Patterns (ODPs)[4], and reusable query templates, to enable more generalizable event pattern descriptions.

Social media streams are becoming an increasingly important asset in achieving situation awareness in many domains and will be also be leveraged in this

[3] http://drools.jboss.org/drools-fusion
[4] http://ontologydesignpatterns.org/

work. Social media streams require systems that can handle incomplete, unreliable, incorrect, and even contradictory data. To handle this it should be possible to employ external applications, e.g., for natural language processing.

The novelty of this approach, compared with traditional CEP, lies in bringing in semantics for the explicit use in decision support, and in taking advantage of Linked Data concepts to abstract heterogeneous data streams. Complex events will be described using a vocabulary that supports layered abstractions of events. By automating various abstraction steps in tractable fashion, the goal is to support decision makers, e.g., by decreasing cognitive workload, assisting event detection, and increase situation awareness.

4 Research Methodology and Approach

Part of the research work involves developing a framework for experimenting with streams, and enabling different RSP engines to be plugged in. This approach is similar to the examples described in Section 2.2. The framework will make it possible to experiment with and benchmark different engines, vocabularies, and queries with minimum overhead. The framework will include adapters for various types of live and recorded streams and support generation of new streams from queries, which will enable the development of various types of stream decorations.

Declarative descriptions will be used to simplify the creation and maintenance of queries, and complex events, and to make some changes to queries possible with minimal user intervention. Appropriate ODPs and query templates will have to be developed. Query templates may require an extension of SPIN[5] (or similar).

5 Preliminary or Intermediate Results

We have previously proposed a number of approaches to representing event object boundaries in various types of RDF streams [9]. While none of the approaches have been evaluated in use, it was shown that if events are described by single RDF graphs boundaries are manageable. In the context of abstracted event objects, consisting of multiple hierarchically ordered graphs, the boundary issues become more complex. Supporting a general view of aggregated and composite event objects will require both a suitable vocabulary, and a clear standard for how RDF streams are represented and communicated.

In [21] we presented a vocabulary that can be used to structure, integrate, and interchange events, regardless of the underlying vocabulary. The model supports multiple time-related parameters, e.g., sampling time, time of entry in the stream, and time of arrival. It also supports the use of payloads, and the encapsulation of event objects. Another important aspect of this vocabulary is that is was defined with querying ability in mind.

[5] http://spinrdf.org/

The requirements and challenges for social media monitoring have also been discussed [10]. In the paper we identified a set of requirements against the current state-of-the-art RSP engines, and highlighted some of the strengths and weaknesses of the systems. While a number of the challenges can be addressed by providing various stream decoration techniques, others are more difficult to address. Only INSTANS is designed to make timestamps available as triples, meaning that it would be difficult for the other engines to support multiple timestamps. The fact that timestamps are not available as triples also creates difficulties with regard to how timestamps should be handled in the case of complex events.

6 Evaluation Plan

The hypothesis of this thesis work will be evaluated within the project *Visual Analytics for Sense-making in CRiminal Intelligence analysis* (VALCRI). The project will enable us to develop and test models for representing complex events, as well as various ways of expressing generalized event patterns, in a real-world setting. The project officially started in late May 2014, and preliminary data and data stream logs have recently been made available.

A survey regarding sate-of-the-art detection of criminal activities will be used as a reference for developing event patterns, query templates, and ODPs. The event patterns and ODPs will be developed in cooperation with the Pacific Northwest National Laboratory. In particular, it will be important to develop event patterns for detecting events from low-frequency signals.

The event patterns will (hopefully) be tested in cooperation with domain experts, acting both as users of the systems and as validators of the generated abstractions. Additionally, validation may be possible based on the data logs themselves, as they contain anonymized records of criminal reports.

7 Reflections

This thesis work aims to develop techniques to assist decision makers in dealing with systems involving large scale heterogeneous real-time data streams by creating high-level abstractions of events. Abstractions in decision making situations can reduce problem complexity and cognitive workload, improve situation awareness, mitigate decision biases, and help to create predictions of future states.

It is clear that streaming data is becoming increasingly important in many domains, but for Semantic CEP to become effective on a broad scale community consensus is required, and there is a need for establishing standard vocabularies, query languages, and stream formats.

When producing declarative descriptions of event patterns the degree to which they can be made scalable is at present not known. All patterns may

not be expressible using the available RSP engines and corresponding query languages, e.g., if windows need to be expressed in the past, or a stream has to be referenced with different windows in the same query.

The representation of complex events risks greatly affecting streaming performance of the engines, and the querying of complex events may introduce considerable overhead and memory problems. Current state-of-the-art RSP engines are still not fully mature, meaning that they are still quite limited in the number of streams and queries they can handle efficiently. One possible scenario is therefore that the systems will only be able to execute a very limited amount of the more complex event pattern queries, or that the queries will result considerable delays.

8 Acknowledgement

The thesis work is financed by *Visual Analytics for Sense-making in CRiminal Intelligence analysis* (VALCRI), *Semantic Technologies for Decision Support* (STeDS), and *Center for Industrial Information Technology* (CENIIT).

References

1. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: A Unified Language for Event Processing and Stream Reasoning. In: Proceedings of the 20th International Conference on World Wide Web (2011)
2. Arasu, A., Babu, S., Widom, J.: The CQL Continuous Query Language: Semantic Foundations and Query Execution. The VLDB Journal 15(2), 121–142 (2006)
3. Balduini, M., Della Valle, E., Dell'Aglio, D., Tsytsarau, M., Palpanas, T., Confalonieri, C.: Social Listening of City Scale Events Using the Streaming Linked Data Framework. In: Proceedings of the 12th International Semantic Web Conference, vol. 8219, pp. 1–16. Springer Berlin Heidelberg (October 2013)
4. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: Querying RDF streams with C-SPARQL. SIGMOD Record 39(1), 20–26 (2010)
5. Blomqvist, E.: The Use of Semantic Web Technologies for Decision Support - A Survey. To appear in: Semantic Web Journal. IOS Press (2012)
6. Druzdzel, M.J., Flynn, R.R.: Encyclopedia of library and information science. In: Bates, M.J., Maack, M.N. (eds.) Encyclopedia of Library and Information Science, chap. Decision Support Systems. Taylor & Francis, Inc., New York, USA, 3 edn. (February 2010)
7. Goodwin, J.C., Russomanno, D.J.: Ontology integration within a service-oriented architecture for expert system applications using sensor networks. Expert Systems 26(5), 409–432 (2009)
8. Gutierrez, C., Hurtado, C.A., Vaisman, A.: Introducing Time into RDF. IEEE Transactions on Knowledge and Data Engineering 19(2), 207–218 (2007)
9. Keskisärkkä, R., Blomqvist, E.: Event Object Boundaries in RDF Streams: A Position Paper. In: ISWC 2013 Workshop: Proceedings of OrdRing 2013 - 2nd International Workshop on Ordering and Reasoning. CEUR workshop proceedings, Sydney, Australia (October 2013)

10. Keskisärkkä, R., Blomqvist, E.: Semantic Complex Event Processing for Social Media Monitoring - A Survey. In: Proceedings of Social Media and Linked Data for Emergency Response (SMILE) Co-located with the 10th Extended Semantic Web Conference. CEUR workshop proceedings, Montpellier, France (May 2013)
11. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. pp. 591–600. ACM, New York, NY, USA (2010)
12. Lampos, V., De Bie, T., Cristianini, N.: Flu detector: tracking epidemics on twitter. In: Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III. pp. 599–602. Springer-Verlag, Berlin, Heidelberg (2010)
13. Le-Phuoc, D., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A Native and Adaptive Approach for Unified Processing of Linked Streams and Linked Data. In: Proceedings of the 10th International Conference on the Semantic Web. pp. 370–388 (2011)
14. Le-Phuoc, D., Parreira, J.X., Hausenblas, M., Hauswirth, M.: Unifying Stream Data and Linked Open Data. Tech. rep., DERI (2010)
15. Liu, D., Pedrinaci, C., Domingue, J.: A Framework for Feeding Linked Data to Complex Event Processing Engines. In: ISWC 2010 Workshop: The 1st International Workshop on Consuming Linked Data (COLD). vol. 665. CEUR workshop proceedings, Shanghai, China (2010)
16. Luckham, D.C.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2001)
17. Marakas, G.: Decision support systems in the 21st century. Prentice Hall, Upper Saddle River, NJ, 2 edn. (2003)
18. Morris, G.W., Thomas, D.B., Luk, W.: FPGA accelerated low-latency market data feed processing. In: Proceedings of the 17th IEEE Symposium on High Performance Interconnects (2009)
19. Quoc, H.N.M., Serrano, M., Le-Phuoc, D., Hauswirth, M.: Super Stream Collider– Linked Stream Mashups for Everyone. In: Proceedings of the Semantic Web Challenge Co-located with the 11th International Semantic Web Conference. Boston, MA, USA (November 2012)
20. Rinne, M., Abdullah, H., Törmä, S., Nuutila, E.: Processing Heterogeneous RDF Events with Standing SPARQL Update Rules. In: On the Move to Meaningful Internet Systems: OTM 2012. Lecture Notes in Computer Science, vol. 7566, pp. 797–806. Springer Berlin Heidelberg (2012)
21. Rinne, M., Blomqvist, E., Keskisärkkä, R., Nuutila, E.: Event Processing in RDF. In: ISWC 2013 Workshop: Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns (WOP2013). CEUR workshop proceedings, Sydney, Australia (October 2013)
22. Sequeda, J.F., Corcho, O.: Linked Stream Data: A Position Paper. In: ISWC 2009 Workshop: Proceedings of the 2nd International Workshop on Semantic Sensor Networks (2009)
23. Slagh, C.L.: Managing Chaos, 140 Characters at a Time: How the Usage of Social Media in the 2010 Haiti Crisis Enhanced Disaster Relief. Master's thesis, Georgetown University, Washington, DC (2010)
24. Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. SIGMOD Record 34(4), 42–47 (2005)
25. Teevan, J., Ramage, D., Morris, M.R.: #Twittersearch: a comparison of microblog search and web search. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 35–44. ACM, New York, NY, USA (2011)