Event Matching using Semantic and Spatial Memories

Majed Ayyad
IT Department (DISI),
University of Trento,
Via Sommarive 14, Trento, I-38123
ayyad@disi.unitn.it

Abstract. We address the problem of real-time matching and correlation of events which are detected and reported by humans. As in Twitter, facebook, blogs and phone calls, the stream of reported events are unstructured and require intensive manual processing. The plethora of events and their different types need a flexible model and a representation language that allows us to encode them for online processing. Current approaches in complex event processing and stream reasoning focus on temporal relationships between composite events and usually refer to pre-defined sensor locations. We propose a methodology and a computational framework for matching and correlating atomic and complex events which have no pre-defined schemas based on their content. Matching evaluation on real events show significant improvement compared to the manual matching process.

1 Motivation and Problem

In recent years a special attention was given to streamed events and stream reasoning [1] [2][13]. A special type of noisy data streamed for real-time reasoning are events which are detected and reported by humans to actionable knowledge bases in multi-tier responding agencies through different services such as Twitter, facebook, phone calls, Microblogs and other similar sources. A common example on this scenario is the stream of incoming phone calls to the operation room of civil police as depicted in Fig. 1. In a standard operation room, operators only register incoming calls, where a second tier of commanders evaluate these calls, support them, if possible, with other information probed from news, blogs and web pages before taking any actions. The second tier is only interested with events that are valid for processing in a time-window. For every new event, they continuously evaluate it against all events in the past time-window in order to find similar clusters of events.

The general main two continuous queries that could be registered on the stream of calls are : **Query 1.** "Compare each incoming event with all previous events logged during the last 5 minutes, then cluster similar events before taking any decision". For the example given in Fig. 1, the query could be translated to "Are these three events the same?". **Query 2.** " Compare each incoming event with all previous events logged during the last hour… Then predict potential new events ".

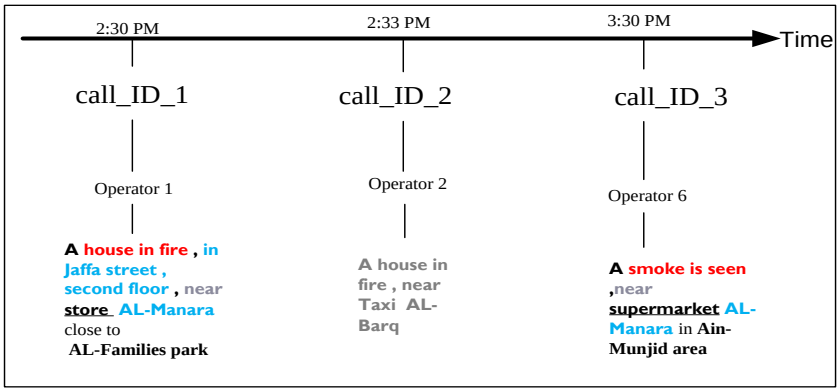

Fig. 1. Calls to the police operation room

To generalize the scenario, and given a stream of events $\{e_1,...e_n\}$ where the structure of e_i, motivated by Davidson convention[4], is of the format $\exists e$ (**Event**(e) $\wedge$ **Agent**(e; an agent) $\wedge$ **Recipient** (e; a recipient) **Time**(e; a time) $\wedge$ **Place**(e; a location) $\wedge$ **Instrument** (e; an instrument)). This format which is illustrated in Fig. 2 also serves as the upper Ontology for events

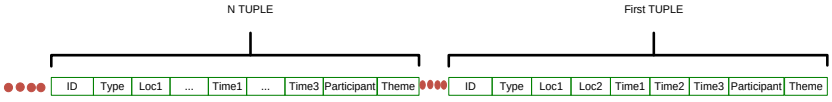

Fig. 2. Stream of non-equal event tuples

The main questions to be answered are :

1. Given a set of atomic events find the similarity between these events in real-time. Similarity is computed as 3-tuples $< e_1, e_2, R >$, where R is expressed as equivalence $(\equiv)$, partially-matched $(\subset)$, and mismatch $(\perp)$.
2. Given a set of occurring events $\{e_1,...e_n\}$ and other historical occurrences, find or infer what pattern of events is occurring.

2 State of the Art

Many approaches followed a content based event matching using different methods which could be summarized as follows **(a) Information retrieval**: [5] use information retrieval techniques for computing events similarity where the event context is treated as a document and the tuple attribute values correspond to document terms. A similar approach was used by The Entity Name System (ENS) [6][7]. **(b) Machine-learning algorithms:** [8] uses machine-learning algorithms to classify events using three groups of features: statistical features, keyword features and word context features. [8] demonstrated that through event mining, it is possible to detect the location and time

for earthquake events by exploiting the real-time nature characteristic of Twitter. The main disadvantage of this method is the need for a large number of events for training, but once learned this method could be used to create models for events correlation.**(c) Predicate-based matching:** The content-based event matching problem was intensively studied in publish-subscribe infrastructure. Where an event to satisfy a subscription, every predicate in the subscription should be matched by some pair in the event [8]. The main disadvantage of predicate-based matching is that predicates should be pre-defined in advance. **(d) Pattern matching (Rete):** The Rete algorithm [9], originally used for production rule systems, is an efficient solution to the facts-rules pattern matching problem. The basic Rete algorithm was extended to accommodate for temporal operators [10][11]. Our approach learns from rete network, but instead of building a network from rules, we build a network from the Ontology and spatial locations.

3 Proposed Approach and Methodology

Our methodology to match and correlate events is based on the content of these events. The methodology approaches the problem from a representational as well as a computational viewpoint as shown on Fig 3. The framework consists of the following components :

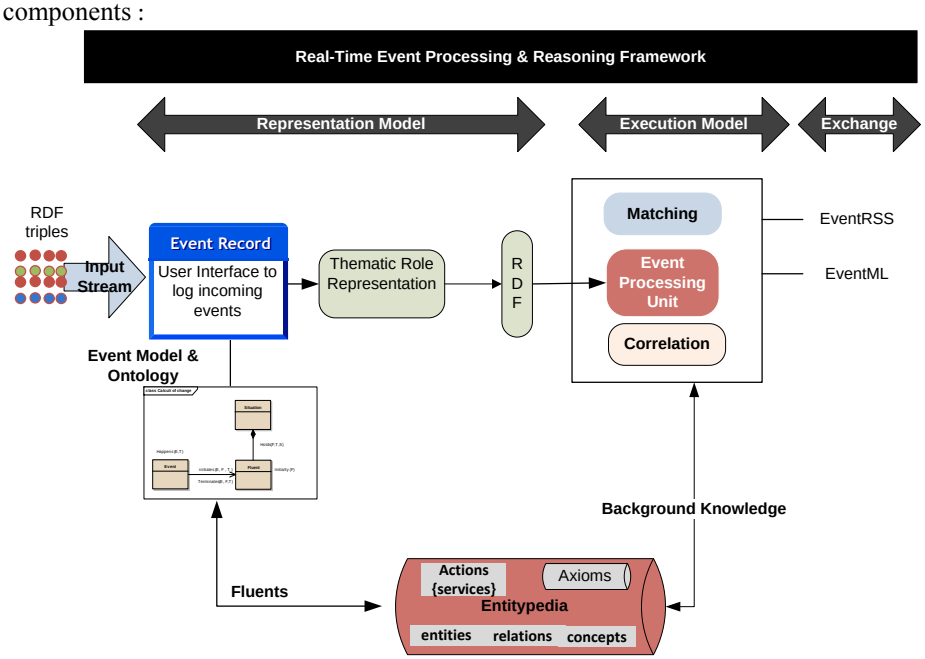

Fig. 3. Real-time event processing framework

(a) **Event Detection and Logging – Controlled GUI :** The first stage, after event detection, starts by logging the event using a controlled natural language

graphical user interface designed to capture temporal and space properties of an event based on pre-defined ontology and model. To build such a model we depend on analyzing the linguistic, Ontological and semantic properties of events and treated events as 4 dimension entities.

(b) Thematic Role Model Builder : The thematic role model aims at representing each event in a form that allows us to correlate and match using the thematic roles of events. Each thematic roles describes the "mode of participation" in an event for each argument of a predicate [3].

(c) **Event Model : Semantic and Spatial Graph representation:** We use two graphs called semantic and spatial memories that are appropriate to perform semantic matching and spatial reasoning about the streamed events. Semantic and spatial memories are built from a central knowledge base of linked entities called Entitypedia[1]. we propose typed composite graphs with inheritance and containment to specify the event structures. After building the network, events asserted from the stream are used to activate these memories at runtime

(d) **Event Query Language :** Clusters of events could be viewed at different granularity based on the typed graphs and their containment relationships.

4 Initial Results and Conclusions

At this stage, we have collected a sufficient number of entities and event types. We collected entities of different types (person, organization and location) from real-life databases. So far we analyzed the meta-data and attributes used by 11 municipalities, 3 Ministries, and two private sector organizations to identify the main entities, their attributes and their instances. We collected 4,358,569 from one country. We designed a preliminary user interface based on the event upper Ontology. For relations between locations, we use the region connection RCC8[12] for qualitative spatial representation and reasoning. For the matching algorithms, we took all the locations in one city and built the RCC8 relationships between these locations. A proof-of-concept prototype for the matching problem was implemented and tested. The initial results show the ability of the system to match hundreds of events efficiently. The set of events that the system couldn't match are collected in a conflict memory. The efficiency of the matching algorithm depends on the number of entities used to build the event networks.

To evaluate the performance of the classification algorithm, we are interested in the algorithm's ability to correctly predict or separate the classes of matched events, partially matched events or non-matched events. To calculate precision and recall we need a ground truth dataset. This data set is under development from multiple sources. During the last six months, events are logged manually on the system from phone

[1] http://entitypedia.org/

calls and two other online news. Crowdsourcing annotations will be used to label events. Disagreement between annotators on event types, spatial and temporal relationships will be evaluated to enhance the parameters of the algorithm.

5 Remaining Work

Still we are working on the optimization of the matching algorithm, specially how to apply different strategies when new token is passed to the event network . Techniques to validate the event ontology and locations path consistency is under consideration. The query language for event matching and correlation with different operators so the end user can be able to examine and fine-tune the obtained results.

References

1. S. Ceri, E. Della Valle, F. van Harmelen and H. Stuckenschmidt, 2010. It's a Streaming World! Reasoning upon Rapidly Changing Information November/December 2009 (vol. 24 no. 6) pp. 83-89
2. Anicic.D and Fodor.P, Rudolph.S, Stojanovic.N.2011.EP-SPARQL: a unified language for event processing and stream reasoning, Proceedings of the 20th international conference on World wide web, March 28-April 01, 2011, Hyderabad, India
3. Carlson, Greg N.1998, Thematic Roles and the Individuation of Events, In Events and Grammar, Vol. 70 (1998), pp. 35-52 Key: citeulike:3137321
4. Davidson,D.1985.The individuation of events, in Davidson (1985) , p 179
5. Kwon,Y.,Lee, W.Y.,Balazinska, M. and Xu, G.2008. Clustering Events on Streams Using Complex Context Information", in Proc. ICDM Workshops, 2008, pp.238-247
6. Bouquet,P.,Stoermer,H. and Bazzanella,B. 2008. An Entity Naming System for the Semantic Web. . In Proceedings of the 5th European Semantic Web Conference (ESWC2008), LNCS, 2008
7. Pernici,B.,Stoermer,H.,Rassadko,N., and Vaidya,N.2010. Feature-Based Entity Matching: The FBEM Model, Implementation, Evaluation. (B. Pernici, Ed.)Lecture Notes in Computer Science Advanced Information Systems Engineering, 6051, 180-193. Springer Berlin Heidelberg. Retrieved from http://www.springerlink.com/content/t784745m2841n52j/
8. Sakaki,T.,Okazaki,M. Matsuo,Y.2010. Earthquake twitter users: Real-time event detection by social sensors. In WWW
9. Forgy, C.L.: Rete: A fast algorithm for the many pattern/many object pattern match problem. Arti_cial Intelligence 19, 17{37 (1982)
10. Berstel, B.: Extending the RETE Algorithm for Event Management. In: Proc. Of 9th Int. Symp. on Temporal Representation and Reasoning (TIME'02). pp. 49{51.IEEE Computer Society (2002)
11. Walzer, K., Breddin, T., Groch, M.: Relative temporal constraints in the Rete algorithm for complex event detection. In: Proc. of 2nd Int. Conf. on Distributed Event-Based Systems. pp. 147{155. DEBS '08, ACM (2008)
12. D. A. Randell, Z. Cui, A. G. Cohn. A Spatial Logic Based on Regions and Connection. In 3rd International Conference on Knowledge Representation and Reasoning (KR 1992), pages 165{176. Morgan Kaufmann, 1992
13. Barbieri,D.,Braga,F.,Ceri,F.,Valle,S., Grossniklaus,M. 2010. Querying RDF Streams with C-SPARQL. ACM SIGMOD Record, 39(1), 20-26. Citeseer. Retrieved from ttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.172.9010