Thingsonomy: Tackling Variety in Internet of Things Events

Souleiman Hasan and Edward Curry, *Insight Centre for Data Analytics at the National University of Ireland, Galway*

Abstract—The Internet of Things (IoT) will connect billions of devices to the Internet and create a large-scale dynamic and open environment with high heterogeneity. Application developers and users need to be abstracted from IoT infrastructure via scalable middleware to assure the rapid adoption of IoT applications. Event processing systems have the potential to contribute in filling the gap between the IoT infrastructure and applications layers. Event processing follows a decoupled model of interaction in space, time, and synchronization. However, the dimension of semantic coupling still exists and poses a challenge to scalability in highly semantically heterogeneous and dynamic environments such as the IoT. In this paper we describe an approach based on loosely coupled producers and consumers enabled with approximate semantic matching of events. We emphasize a practitioner perspective to IoT architectures for building software that can tackle heterogeneity of IoT events.

Index Terms—Internet of Things, distributed applications, event processing, semantic normalization, IoT architecture.

1 Introduction

The Internet of Things (IoT) builds upon the success story of Internet technologies to connect physical objects, or things, to the Internet and enable a plethora of applications such as assisted driving, augmented reality, smart and comfortable homes, etc [1]. A basic requirement to realize the IoT is an infrastructure of communication solutions and interoperability standards such as the Constrained Application Protocol (CoAP) by the Internet Engineering Task Force (IETF) [1]. There is also a need for middleware that can abstract the application developers from the underlying technologies which is crucial to the adoption and evolution of IoT applications [1].

Event-based technology has played an important role in the middleware space to enable scalable software architecture based on its loosely coupled model of interaction. Nevertheless, event-based systems assume a high level of semantic agreement between event producers and consumers which is challenging for largely heterogeneous environments such as smart cities due to the difficulties to establish common semantic agreements. Current approaches use granular semantic models such as ontologies but such models are time consuming to build and agree upon and thus limit scalability.

This paper extends the event-based architecture to encompass the semantic normalization functionality needed in IoT. It guides practitioners to build IoT applications where exchanged events convey semantics and at the same time frees parties from rigid agreements. This is based on: (1) a semantic model based on terms statistical co-occurence in large textual corpora such as Wikipedia, (2) thematic tagging of events and subscriptions, and (3) an approximate probabilistic matcher of events.

2 The Internet of Things and Event-based Systems

From a high-level architectural perspective IoT can be divided into three tiers [1]:

1) **Sensing and communication** technologies which form the basic infrastructure for IoT to map the world of things into the world of computationally processable information. Radio-frequency Identification (RFID) plays a key role within this tier where RFID tags are attached to real world things and RFID readers are responsible for instrumenting their information into the Internet. Communication and networking standards such as the IPv6 over Low power Wireless Personal Area Networks (6LoWPAN) and the CoAP protocols [1] serve this layer of IoT.
2) **Middleware layer** which encompasses common functionalities and abstracts application developers and users from IoT infrastructure details. Among the technologies to contribute to this layer are Service-Oriented Architectures (SOA) [1] and the Message-Oriented Middleware (MOM). Event processing systems are a more generic version of MOM which support functionalities such as early filtering of events, spatio-temporal correlation, sequencing, event enrichment, event aggregation, and complex event processing.
3) **Application layer** which builds upon the middleware to provide direct and potentially domain specific benefits to users. IoT promises new domains of applications in transportation and logistics, healthcare, smart environments, analytics, personal and social media, etc.

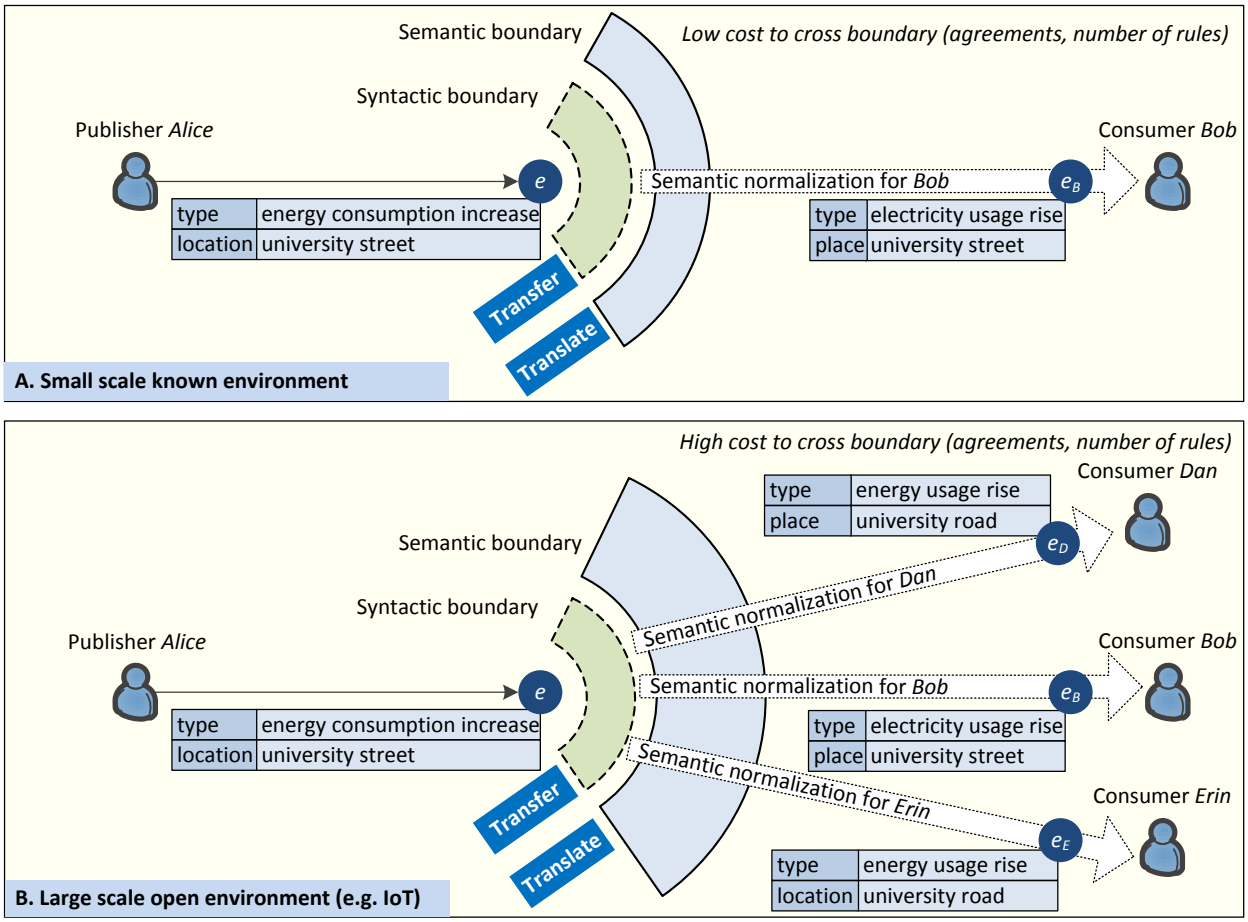

Fig. 1: Boundaries to event exchange in (A) a small scale known environment, and (B) a large-scale open environment such as the Internet of Things.

3 Motivation

Bob works in the town hall planning department of a smart city. Bob is interested in finding the energy usage of street lights during peak electricity usage in different areas. Such information can be detected using an expression of an Event Processing Language (EPL) such as Esper's language [2] as follows:

every a=StreetLightsEvents(
a.type= 'energy consumption event'
and a.area.consumptionPeak='true')

While the sources of required information are available from the street lights, the semantics of the events differ from one area to another due to different sensors manufacturers. For instance, events contain terms such as *'energy consumption'*, and *'electricity usage'* to refer to the same thing. The scenario requires a large set of rules with high definition and maintenance costs in order to cover events semantic heterogeneity.

4 Cross Boundary Information Exchange

In a system of systems such as the Internet of Things, information items such as events need to cross system boundaries to enable cooperation. Carlile [3] recognizes two main levels of boundaries that may exist in a given knowledge exchange scenario:

- *Syntactic boundary* affects the basic knowledge *transfer* mechanism between participants. In a broad sense, it is concerned with data formats, participants interaction time, and addressing which are expected to exist in most event-based environments as shown in Figure 1.
- *Semantic boundary* starts to appear when new event sources or consumers make some meanings unclear or ambiguous. Semantic boundaries are inherent in large-scale, open and heterogeneous environments such as the IoT as shown in Figure 1. Thus, *translating* heterogeneous information items into a common meaning model that a developer can work with is crucial. We call this *semantic normalization* and it effectively means crossing the semantic boundaries between systems.

The Internet of Things requires an open mode of information exchange in which systems boundaries have to be crossed frequently. This puts openness as an inevitable requirement that needs to be met by technologies used to realize the Internet of Things. Event-based

CURRENT APPROACHES TO SEMANTIC NORMALIZATION are shown in Table 1. In the *content-based approach*, event sources and consumers use the same event types, attributes and values as assumed in traditional content-based publish/subscribe systems such as SIENA [4]. The approach has high semantic coupling between parties and works well in environments with a low level of data heterogeneity. In the *concept-based approach*, participants can use different terms and still expect event matchers to match them properly thanks to an explicit knowledge representation that encodes semantic relationships between terms. Examples of knowledge representations are thesauri and ontologies as in S-TOPSS [5] and semantic pub/sub [6]. Building such knowledge representations is a time consuming process.

TABLE 1: Approaches to Semantic Normalization [7]

	Content-based [4]	**Concept-based** [5], [6]	**Approximate Semantic Event Processing** [8], [9]	**Thematic Event Processing** [7]
Matching	exact string matching	Boolean semantic matching	approximate semantic matching	approximate semantic matching
Semantic coupling	term-level full agreement	concept-level shared agreement	loose agreement	loose agreement
Semantics	not explicit	top-down ontology-based	statistical distributional semantics	statistical distributional semantics
Domain specificity cost	defining a large number of domain rules	defining a domain-specific ontology	indexing a domain-specific corpus	parametrizing the vector space of an open domain corpus
Effectiveness (F_1Score)	100%	depends on the domains and number of concept models	depends on the corpus	depends on the corpus and the themes tags. Outperforms non-thematic approximate approach
Cost	defining a large number of rules and establishing shared agreement on terms	establishing shared agreement on ontologies	minimal agreement on a large textual corpus	minimal agreement on a large textual corpus and associating good themes tags
Efficiency (throughput)	high	medium to high	medium to high	medium to high

Freitas et al. proposed an approximate query processing approach for databases based on distributional semantics [10]. In our previous work [8], [9], we proposed an *approximate semantic event processing approach* and showed that the model is suitable when participants agree on some event types, attributes, or values while performance decreases significantly with an absolute 100% degree of required approximation.

systems have great potential to contribute to realizing the IoT due to their decoupled nature. Nonetheless, they do not easily cross semantic boundaries due to assumptions of semantic agreements on terms within events and subscriptions.

In the event-based paradigm, event sources fire instantaneous and atomic information items called events. Event consumers use rules or subscriptions to detect events and react to them. Events are the only means of interaction between sources and consumers. This results in decoupling the production and consumption of events and thus increasing scalability by "removing explicit dependencies between the interacting participants" [11].

Event-based systems decouple participants on three dimensions [11]:

- *Space decoupling* suggests that participants do not need to know each other.
- *Time decoupling* means that participants do not need to be active at the same time.
- *Synchronization decoupling* suggests that event producers and consumers are not blocked while producing or consuming events.

The space, time, and synchronization decoupling dimensions of Eugster et al. [11] can be seen to contribute to event transfer across Carlile's syntactic boundaries.

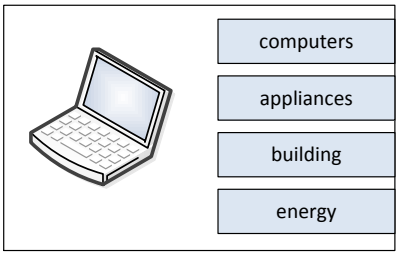

Fig. 2: An example thingsonomy for tagging a device's events.

However, event-based systems can be at the same time tightly coupled by the semantics of exchanged events. Traditional deployments of event systems assume a mutual agreement on event types, attributes, and values to achieve semantic normalization and that forms an explicit dependency between participants. For example, if a smart city event source marks an event with the type *'parking space occupied'*, all event consumers of this event would have to use this exact event type in their rules. A new event consumer to the system cannot use a rule with the term *'garage spot occupied'* to handle such events.

5 Requirements

This paper tackles the following requirements to address event variety in IoT middleware and application layers:

- Low cost for integrating and accessing heterogeneous IoT devices. A main task for integration is the normalization of several heterogeneous data items into common models.
- Effective and near real-time processing of IoT events. Processing middleware should be able to match items of interest with a high detection rate of true positives and negatives and with low latency.

6 Thematic Event Processing and 'Thingsonomies'

Our thematic event processing approach builds on the analogy of the wide spread use of social tagging, or folksonomies [12]. It has been observed that imposing fixed or agreed-upon top-down taxonomies on users to describe web content such as images is unfeasible [12]. Instead, bottom-up and user generated tags called folksonomies are used by users to tag and discover content. Consequently, many social tagging platforms have flourished such as Flicker, Twitter, Delicious, etc.

We suggest associating thematic tags that describe the themes of types, attributes and values and clarify their meanings. We call these tags *thingsonomies* for things and taxonomies. The hypothesis is that associating events and subscriptions with extra tags can improve effectiveness and time efficiency in heterogeneous environments and domain-specific knowledge exchange.

Distributional semantics is based on the hypothesis that similar and related words appear in similar contexts. Distributional models are quite useful for the task of assessing semantic similarity and relatedness between terms. A *semantic measure web service* of Figure 3 is a function that quantifies the similarity/relatedness between two terms and typically has its values in $[0, 1]$. Distributional models can be constructed automatically from statistical co-occurrence of words in a *corpus of documents*. This model is formalized as a *vector space* which provides a computationally efficient framework for calculating similarity scores and represents a good fit for the requirements of loose semantic coupling and real-time performance for event-based IoT.
A widely used example is the distributional Explicit Semantic Analysis semantic measure *esa* constructed from Wikipedia corpus [13]. In a nutshell, Wikipedia-based *esa* builds an index of words based on the Wikipedia articles they appear in, hence *indexing* in Figure 3 [14]. A word becomes a vector of articles and the more common articles between two words exist, the more related the words are. For example, *esa('parking', 'garage') > esa('parking', 'energy')* as the formers appear frequently in common articles. Typically semantic relatedness between a pair of terms is measured using cosine distance between the two vectors representing the two terms. In our thematic model, *esa* measure is parametrized also with the themes tags. Those are used to project the terms vectors to get a more domain-specific meaning vectors and then are passed to the distance function.

Figure 2 shows an example thingsonomy for tagging energy consumption events coming from a laptop.

Figure 3 illustrates the main components of the thematic event processing approach. Thematic events can cross semantic boundaries as: (1) they free users from a priori semantic top-down agreements and thus enable event exchange across such boundaries, and (2) they carry approximations of events meanings composed of payloads and thematic tags which when combined carry less semantic ambiguities. An approximate matcher exploits the associated thematic tags to improve the quality of its uncertain matching.

Step 1 to build the IoT architecture enabled with semantic normalization is to build a semantic model which enables the system to automatically establish relationships between various terms such as *'computer'* vs. *'laptop'*. Our approach adopts a distributional model of semantics based on statistical indexing of a large corpus of textual documents, refer to the sidebar. Such a model

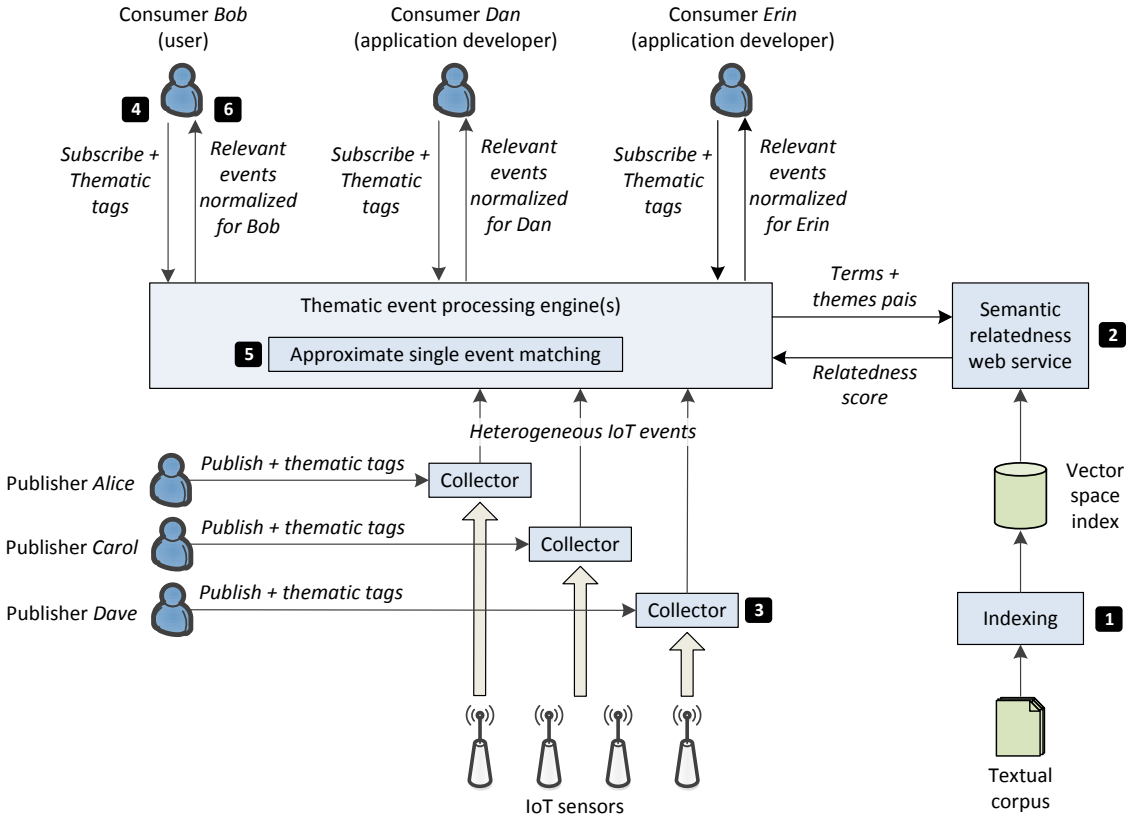

Fig. 3: Architecture for loosely coupled semantic normalization for Internet of Things software.

is easy to build automatically as shown in [14], and the main task for the practitioner is the corpus selection. One can start working with an initial documents corpus, e.g. Wikipedia, and incrementally revise it to suit the use cases.

Step 2 is to avail a semantic relatedness measure based on the built semantic model through a conventional interface such as REST and JSON [14]. For example, a request for relatedness between *'electricity'* and *'energy'* is invoked through API:

```
http://example.com/esa?term1=energy
&term2=electricity
```

with the result being returned as a JSON object as follows:

```
{"relatedness" : 0.154}
```

Such a result makes sense only in comparison with the relatedness of other terms such that *'electricity'* is closer to *'energy'* than to *'office'* for instance.

Step 3 is for publishers to accompany their events with a set of thematic tags at the data collector. Such tags shall represent approximately the domain and meaning of the terms used to describe the event attributes and values. Let an event of an *increased energy consumption* be represented as follows:

{**type**: increased energy consumption event,
measurement unit: kilowatt hour,
device: computer, **office**: room 112}

An example of thematic tags for this event are:

{computer, appliances, building, energy}

Step 4 is for subscribers to associate their subscriptions with thematic tags. We use a language that introduces the *tilde* $\sim$ operator which signifies that the user wants the matcher to match the term used or any term semantically similar to it. A subscription for *increased energy consumption* can be represented as follows:

{**type**= increased energy usage event$\sim$,
device$\sim$= laptop$\sim$, **office**= room 112}

Example thematic tags are:

{power, computers}

Step 5 is the responsibility of the system to normalize events and match them to the suitable subscriptions. The example event and subscription do not use exactly the same terms to describe the type or the device, hence *'energy consumption'* vs. *'energy usage'*, and *'computer'* vs. *'laptop'*. Nevertheless, the event should not be considered as a negative match to the subscription. For this reason, our model employs a probabilistic matcher which uses a measure to estimate semantic similarity and relatedness

EVALUATION of the normalization quality can be achieved by establishing a gold standard set of subscriptions and events of known ground truth of true matchings. For each subscription, the set of relevant events is identified. *Precision* represents the ratio of correctly matched events versus all the matched events. *Recall* represents the ratio of correctly matched events versus all the relevant ones. The effectiveness of the built software can be measured by precision, recall, and a derivative measure that combines both in one number such as the $F_1 Score$. Efficiency can be measured using *event throughput* which represents the amount of processed events per a time unit in the IoT middleware layer from the sensors to the applications.

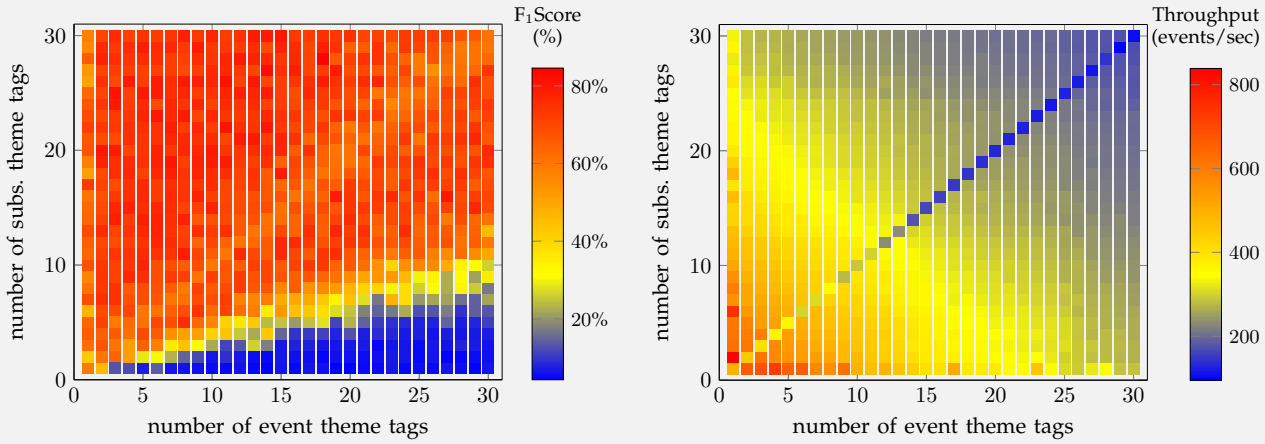

Fig. 4: Evaluating IoT semantic normalization: effectiveness (left) and time efficiency (right) [7].

Test events and subscriptions sets shall be chosen based on the use cases. For example, in [7] we have synthesized a set of around $15,000$ events of up to 10 attribute-value pair per event, and around 100 approximate subscriptions from real world smart city deployments in Europe such as the SmartSantander project [15] which employs a set of sensors to monitor temperature, noise, traffic, parking, and others.
Seed events are expanded into the final set, and ground truth matching and thematic tags were generated. Figure 4 illustrates the resulting effectiveness and efficiency of the approximate matcher working with Wikipedia-based *esa*. Each cell in the figure shows the result that corresponds to a combination of numbers of thematic tags associated with events (the X-axis), and subscriptions (the Y-axis).
Results show that the thematic approach is limited when users can provide only a small number of tags for subscriptions, and when hard real-time deadlines are required. Otherwise, results suggest that the use of less terms to describe events, around $2-7$, and more to describe subscriptions, around $2-15$, can achieve a good matching quality, up to 85%, and throughput, up to 800 events/sec, together with less error rates. That is concentrated in the middle left part of squares in Figure 3 (more red cells).
Results also show that the approach is scalable to highly semantically heterogeneous environments due to the lightweight amount of tagging required and the low number of approximate subscriptions which is about 100 subscriptions. That would *cost* users an equivalence of around $48,000$ exact subscription rules.

between various terms. Functionally, it tries to establish possible mappings between subscription predicates and event tuples. For example, the most probable mapping of previous examples is described as follows:

$\sigma^* =$ {(**type=increased energy consumption event** $\leftrightarrow$ type:increased energy usage event),
(**device**$\sim$ **= laptop**$\sim$ $\leftrightarrow$ device:computer),
(**office = room 112** $\leftrightarrow$ office: room 112)}

Step 6 represents the return of events matching a subscription to its initiator. The matcher establishes probabilistic matching and as a result forwards the normalized event along with an uncertainty value that reflects the amount of semantic normalization that has been conducted all the way from publishers to subscribers.

To evaluate the proposed architecture, a framework conceived from the evaluation of Information Retrieval search engines is used. The framework is built upon the concepts of matching *precision*, *recall*, and $F_1 Score$ along with *throughput* as discussed in the top sidebar.

7 DESIGN CONSIDERATIONS

The degree of approximation is the number of tilde $\sim$ operators used in subscriptions. It can be used to quantify the approximation done by the engine during semantic normalization. The proposed approach works better and needs less tags with lower degrees of approximations as exact string matching can help filter many events. For

example, in some applications several agreements can be assumed such as units of measurements as in smart grids.

Besides, the use of semantic relatedness services instead of exact string comparison is costly from a time performance point of view. Thus, applications with hard real-time deadlines such as some security systems may not be the ideal applications. It could be better to afford the cost of establishing semantic agreements and use a traditional publish/subscribe system rather than leaving semantic approximation to the matcher.

8 Conclusions and Future Work

We have discussed the challenge of building IoT software that overcomes event semantic variety in a loosely coupled manner. We highlighted the practical aspects for building IoT software via thingsonomies for semantic normalization in an event-based middleware. Future work for practitioners is to test the suitability of various corpora with respect to each domain such as energy, traffic, etc. It also includes the use of cloud computing and parallel processing to improve efficiency within applications that have real-time constraints.

Acknowledgments

This work has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787 – 2805, 2010.

[2] EsperTech, "Esper Complex Event Processing Engine," 2014. [Online]. Available: http://esper.codehaus.org/.

[3] P. R. Carlile, "Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries," *Organization science*, vol. 15, no. 5, pp. 555–568, 2004.

[4] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf, "Achieving scalability and expressiveness in an internet-scale event notification service," in *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing*, ser. PODC '00. New York, NY, USA: ACM, 2000, pp. 219–227.

[5] M. Petrovic, I. Burcea, and H.-A. Jacobsen, "S-topss: Semantic toronto publish/subscribe system," in *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, ser. VLDB '03. VLDB Endowment, 2003, pp. 1101–1104.

[6] L. Zeng and H. Lei, "A semantic publish/subscribe system," in *E-Commerce Technology for Dynamic E-Business, 2004. IEEE International Conference on*, Sept 2004, pp. 32–39.

[7] S. Hasan and E. Curry, "Thematic event processing," in *Proceedings of the 15th International Middleware Conference*, ser. Middleware '14. New York, NY, USA: ACM, 2014, pp. 109–120.

[8] S. Hasan, S. O'Riain, and E. Curry, "Approximate semantic matching of heterogeneous events," in *Proc. The 6th ACM International Conference on Distributed Event-Based Systems*, ser. DEBS '12, 2012, pp. 252–263.

[9] S. Hasan and E. Curry, "Approximate semantic matching of events for the internet of things," *ACM Trans. Internet Technol.*, vol. 14, no. 1, pp. 2:1–2:23, Aug. 2014.

[10] A. Freitas, J. G. Oliveira, S. ORiain, E. Curry, and J. C. P. Da Silva, "Querying linked data using semantic relatedness: a vocabulary independent approach," in *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB)*. Springer, 2011, pp. 40–51.

[11] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," *ACM Computing Surveys (CSUR)*, vol. 35, no. 2, pp. 114–131, 2003.

[12] M. Gupta, R. Li, Z. Yin, and J. Han, "Survey on social tagging techniques," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 58–72, 2010.

[13] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.

[14] D. Carvalho, C. Çalli, A. Freitas, and E. Curry, "EasyESA: A low-effort infrastructure for explicit semantic analysis (demonstration paper in proceedings)," in *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)*, 2014.

[15] L. Sanchez, J. A. Galache, V. Gutierrez, J. Hernandez, J. Bernat, A. Gluhak, and T. Garcia, "Smartsantander: The meeting point between future internet research and experimentation and the smart cities," in *Future Network & Mobile Summit (FutureNetw), 2011*. IEEE, 2011, pp. 1–8.

Souleiman Hasan is a PhD student at the Insight Centre for Data Analytics at the National University of Ireland, Galway (formerly DERI). His main research interests include the Internet of Things, semantic web, event processing, and semantic matching. Hasan has a BSc in information technology engineering, major of software engineering and information systems from Damascus University. Contact him at souleiman.hasan@insight-centre.org.

Edward Curry is a research leader at the Insight Centre for Data Analytics at the National University of Ireland, Galway (formerly DERI), and an adjunct lecturer at NUI, Galway. His projects include studies of the Internet of Things, enterprise linked data, energy informatics, semantic information management, and community-based data curation. Curry has a PhD from NUI, Galway. Contact him at ed.curry@insight-centre.org.