Thematic Event Processing

-Research Paper-

Souleiman Hasan
Insight @ NUI Galway
Galway, Ireland
souleiman.hasan@insight-centre.org

Edward Curry
Insight @ NUI Galway
Galway, Ireland
ed.curry@insight-centre.org

ABSTRACT

Event-based systems follow a decoupled mode of interaction between event producers and consumers in space, time, and synchronization to enable scalability within distributed systems. We recognize a fourth dimension of coupling due to the need for mutual agreements on terms that describe event types, attributes, and values. Semantic coupling is challenging in large-scale, open, and heterogeneous environments such as the Internet of Things (IoT). It requires event producers and consumers to agree on event semantics and can limit scalability due to the difficulties in establishing such agreements. In this paper we propose a new *thematic event processing* approach based on enhancing events and subscriptions with terms representing their themes to clarify their domains and meanings in addition to their payload. Experiments conducted using large heterogeneous sets of smart-city and energy management events suggest up to 85% of matching accuracy at a rate of 500 events/sec of throughput. This represents around 15% improvement in accuracy and 150% in throughput over non-thematic approaches. This suggests the viability of thematic event processing to scale to environments such as the IoT.

Categories and Subject Descriptors

C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Distributed systems*

General Terms

Algorithms, Experimentation, Performance, Languages

Keywords

Approximate matching, distributional semantics, event processing, Internet of Things, semantic matching, theme tags, uncertainty

Middleware '14, December 08 - 12 2014, Bordeaux, France
Copyright 2014 ACM 978-1-4503-2785-5/14/12 ...$15.00
http://dx.doi.org/10.1145/2663165.2663335.

1. INTRODUCTION

It is estimated that 50 billion devices will be connected to the Internet by 2020 [21] forming the Internet of Things (IoT). IoT builds upon the success story of Internet technologies to connect physical things to the world wide Internet [3]. IoT projects such as the SmartSantander smart city have already deployed tens of thousands of Internet connected sensors in large cities [23] to monitor temperature, noise, traffic, parking, and others.

IoT is based on an infrastructure of communication standards such as the 6LoWPAN and the CoAP protocols [3]. Higher up the stack, it will enable a plethora of applications including assisted driving, augmented reality, smart homes, etc [3]. In-between, there is a need for middleware to abstract application developers from underlying technologies to facilitate the adoption of IoT applications [3]. Event-based technology has played an important role in the middleware space, specifically in enterprise application integration, and enabling Internet scale distributed systems [7].

Event-based technology is based on a loosely coupled interaction model which supports scalability. Nevertheless, it assumes a high level of semantic agreement between event producers and consumers. Small and controlled environments have: a small number of event sources and types, a low degree of event data heterogeneity, a small number of users, and event consumers who understand the environment. It is possible to establish agreements on event semantics and maintain rules that can cover the heterogeneity of events. However, this will be challenging in large and open environments such as IoT smart cities due to thousands of potential event sources and types, many systems which publish and consume events, and difficulties to establish top-down semantic agreements or maintain all possible rules that can cover all potential events heterogeneity.

We believe that event-based middleware needs to support data management functionalities at such large scales. This paper proposes a thematic event processing model where approximations of events meanings are exchanged to complement attributes and values. The model requires loose semantic agreements between participants and a small amount of subscriptions to cover events heterogeneity while achieving satisfactory matching quality and throughput results.

1.1 Background

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

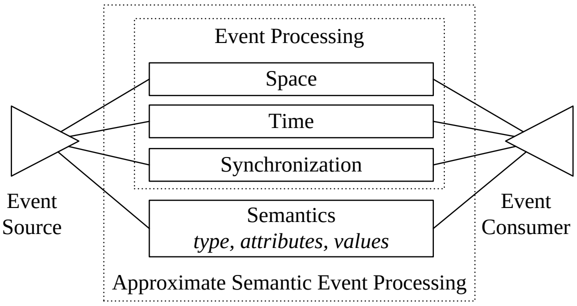

Figure 1: Decoupling dimensions.

between sources and consumers. This results in decoupling the production and consumption of events and thus increases scalability by "removing explicit dependencies between the interacting participants" [9]. Event-based systems decouple participants on three dimensions as shown in Figure 1:

- *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 [9].

Nevertheless, 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 which 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 the event.

The relative importance of each coupling dimension varies with the boundaries which exist in a target system of systems. Carlile [5] 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. 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 2. They form the basics in information communication as originally discussed in Shannon and Weaver's information theory [24]. The space, time, and synchronization decoupling dimensions of Eugster et al. [9] can be seen to contribute to event transfer across this type of boundaries.
- *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 2. Thus, establishing mutual agreements on event semantics (semantic coupling) becomes crucial. This in turn leads to magnifying the problematic nature of semantic coupling which contradicts with the fundamental basis of event systems as decoupled and scalable systems.

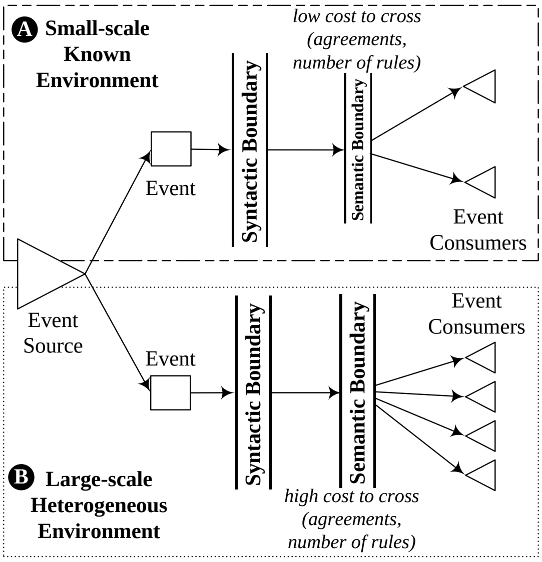
A Small-scale Known Environment
low cost to cross (agreements, number of rules)
Syntactic Boundary
Semantic Boundary
Event
Event Consumers
Event Source
Event
Syntactic Boundary
Semantic Boundary
Event Consumers
B Large-scale Heterogeneous Environment
high cost to cross (agreements, number of rules)

Figure 2: Boundaries to event exchange in (A) a small scale known environment, and (B) a large-scale heterogeneous environment.

Current event-based middleware can still be applied within environments with non-significant semantic boundaries such as enterprise environments. The approach in this paper suggests associating events and subscriptions with additional meaning signposts called themes when semantic boundaries start to appear in open, heterogeneous, and dynamic environments. It forms a compromise between absolute mutual agreements on semantics which constrains scalability and absolute ignorance of semantic agreements which leads to high levels of false event matching.

1.2 Current Approaches

There are 3 main approaches to the semantic coupling problem as shown in Table 1 and discussed in the following.

1.2.1 Content-based Approach

In content-based approaches, event sources and consumers use the same event types, attributes, and values without any extra description of meaning external to the subscriptions and events. This is assumed in traditional content-based publish/subscribe systems such as SIENA [7] where the matcher performs string exact comparison between terms. The approach has high semantic coupling between parties and works well in environments with a low level of data heterogeneity. However, it becomes difficult to scale to more heterogeneous environments due to the effort required to keep the agreement on shared schemata and to develop a large number of subscriptions according to the agreement.

1.2.2 Concept-based Approach

In the concept-based approach, participants can use different terms and still expect matchers to match them properly thanks to an explicit knowledge representation that encodes semantic relationships between terms. Example knowledge representations are thesauri and ontologies as in S-TOPSS [22] and semantic pub/sub [28], as well as programming language type hierarchies as in the type-based

Table 1: Approaches to Semantic Coupling

	Content-based	**Concept-based**	**Approximate Semantic Event Processing**	**Proposed Thematic Event Processing**
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 model based on distributional semantics	statistical model based on 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 a sufficiently large and comprehensive 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, refer to Section 5.3.2

publish/subscribe model [10]. Building such knowledge representations is time consuming and agreements suggest an explicit dependency between parties, not directly but via the conceptual model. Thus, relatively high semantic coupling exists as agreement is needed for each individual concept.

1.2.3 Approximate Semantic Event Processing

In the absence of an agreement on event schema or a conceptual model, participants may loosely agree on topics represented in large corpora of texts. Such corpora can be used to automatically construct distributional models of meaning to derive semantic similarity and relatedness based on terms co-occurrence. Freitas et al. proposed an approximate query processing approach for databases based on distributional semantics and validated it within a natural query scenario over graph databases [11]. In previous works [16] [17], we proposed and discussed the details for an approximate semantic event processing approach based on approximate probabilistic matching and semantic similarity. Our previous work in [16] tackled semantic coupling within domain-agnostic environments. Experiments showed that the model is suitable when participants agree on some event types, attributes, or values while performance decreases significantly when an absolute 100% degree of approximation is required.

This paper introduces the concept of event themes which is a pragmatic compromise to semantic coupling. We believe the introduction of themes is a critical step that makes the approximate semantic approach a viable technique within real-world use cases. Our work in [16] stresses the importance of approximate matching, but this paper stresses the exchange of meanings, e.g. attributes/values + tags+ background distributional vector space, as first class citizens in event systems to loosen semantic coupling.

The rest of this paper is organized as follows: The proposed approximate thematic event processing model and the contributions are discussed in Section 2. Model instantiation for structured attribute-value events is detailed in Sections 3 and 4. Section 5 details the evaluation methodology and results. Related work is discussed in Section 6. Section 7 discusses future work and concludes the paper.

2. THEMATIC EVENT PROCESSING

The generic thematic event processing model is motivated and discussed throughout the following subsections.

2.1 Motivational Scenario

Alice works in the town hall planning department of a smart city. Alice is interested in finding the energy usage of street lights during peak electricity usage in their areas. Such information can be detected using an Event Processing Language (EPL) such as Esper's[1] as follows:

```
pattern [ every a=StreetLightsEvents(
    a.type= ‘energy consumption event’
    and a.area.consumptionPeak=‘true’)]
```

While the sources of required information are available, 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 IT department requires a large set of rules such as the one above with all possible variations of semantics in order to cover the events semantic heterogeneity. Definition and maintenance of such rules requires significant time and effort.

2.2 Requirements and Questions

The main requirement tackled in this paper is to reduce the effort to describe the meaning of events and subscrip-

[1] http://esper.codehaus.org/

tions. The research questions stemming from this requirement are: (1) how to achieve semantically loosely coupled event exchange? and (2) how to effectively and efficiently match exchanged events in the environment?.

2.3 Proposed Approach

The approach in this paper builds on the analogy with the wide spread use of social tagging, or folksonomies, in knowledge discovery [14]. Web 2.0 forms a large-scale environment where users are decoupled and distributed. It has been observed that imposing fixed or agreed-upon top-down taxonomies on users to describe web content such as images is unfeasible [14]. Instead, bottom-up and user generated tags called folksonomies are used by users to tag and discover content. For example Xu et al. [27] showed that using folksonomies for information retrieval significantly improves search quality. Consequently, many social tagging platforms have flourished such as Flicker, Twitter, Delicious, etc.

The proposed approach suggests associating representative terms that describe the themes of types, attributes and values and clarify their meanings as shown in Figure 3. The hypothesis is that associating events and subscriptions with extra information that better describes their meanings can improve effectiveness and time efficiency in heterogeneous environments and domain-specific knowledge exchange. Thematic events can more easily cross semantic boundaries as: (1) they free users from needing a prior semantic top-down agreements and thus enable event exchange across such boundaries, and (2) they carry approximations of events meanings composed of payloads and theme tags which when combined carry less semantic ambiguities. An approximate matcher exploits the associated themes tags to improve the quality of its uncertain matching of events and subscriptions.

This generic architecture applies to various types of event payloads. For example, an event payload can be an image and its theme is a set of tags describing its content like {*'girl', 'football', 'outdoor'*}. A subscription can be an image too associated with a set of tags such as {*'female', 'ball', 'play', 'nature'*}. The approximate matcher performs an uncertain matching on images based on their pixels and other intrinsic image features. It also exploits the tags associated with the event and the subscription to parametrize its matching algorithm and improve its matching quality. For instance it weighs up some object recognition candidates more like *'girl'* versus *'boy'* in the event image. Event sources and consumers can either (1) agree on the use of representative terms when agreement is possible and thus having lightweight *loose coupling*, or (2) freely use representative terms in open environments when agreement is not possible, thus having *no coupling*.

This paper instantiates a generic thematic event processing model for structured attribute-value events and subscriptions. The attribute-value model is simple, widely used, and may be used to convey other models. Theme tags are exchanged with the events and used by the matcher to more accurately filter a distributional representation of terms in a vector space as discussed in Sections 3 and 4.

2.4 Contributions

The contributions of this paper are:

- A thematic event processing model to address semantic coupling in heterogeneous and domain-specific event exchange environments effectively and efficiently.

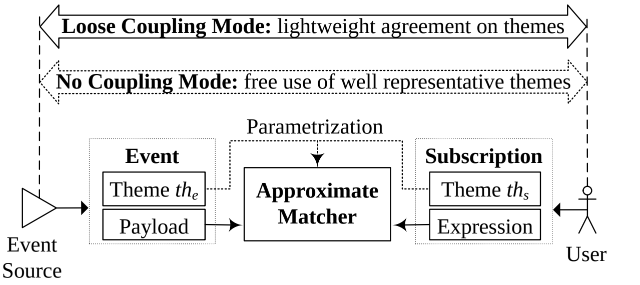

Figure 3: Thematic event processing.

- A formal framework for structured events based on thematic projection in parametric vector space.
- An evaluation framework based on synthetic event loads and approximate subscriptions from real world IoT deployments and domain-specific thesauri.

3. MODEL INSTANTIATION

The main elements of the model instantiation are illustrated in Figure 4. Let an event of *increased energy consumption* be represented as follows:

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

In the thematic model, this event is accompanied with a set of key terms that represent approximately the domain and meaning of the event attributes and values. We call these terms the *event theme tags*. An example of terms for the above event are:

{energy, appliances, building}

Similarly, subscriptions are associated with *subscription theme tags*. The proposed model language 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 theme tags for this subscription are:

{power, computers}

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 an approximate probabilistic semantic matcher which uses a measure to estimate semantic similarity and relatedness between various terms. Functionally, it tries to establish the top-1 or top-k possible mappings between subscription predicates and event tuples along with probability spaces of each predicate-to-tuple and of the overall mapping. For example, the most probable mapping of the previous examples, or top-1 mapping, is described as follows:

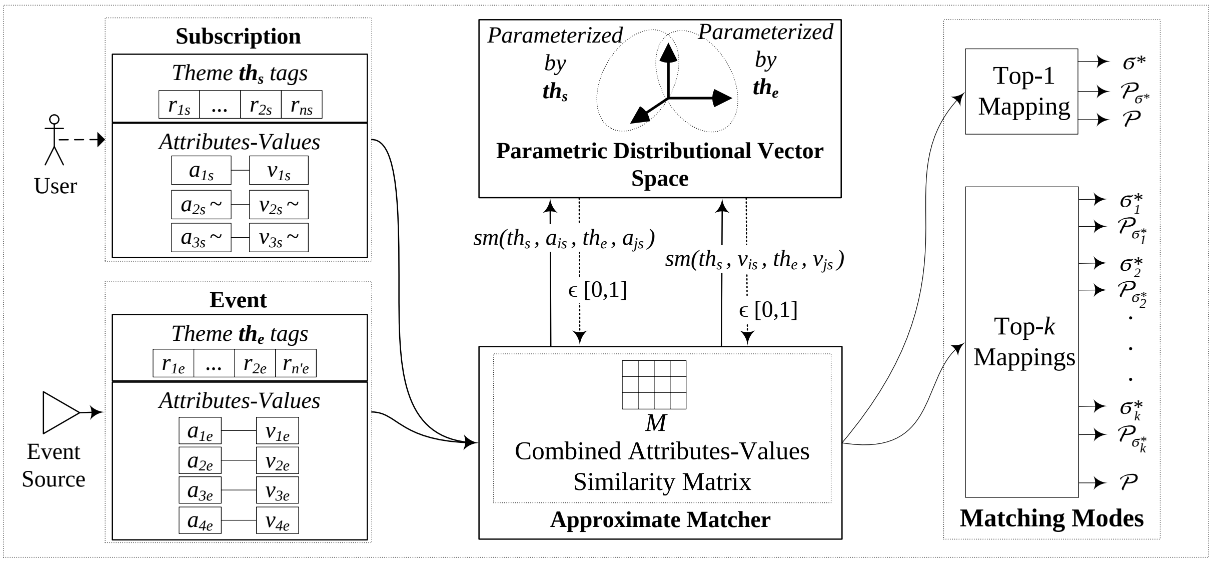

Figure 4: Thematic event matching model.

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

The approximate matcher uses a semantic measure to estimate semantic similarity and relatedness between each pair of attributes or values from the subscription and the event. The matcher then combines that in a similarity matrix that encodes similarity between all possible pairs of subscription predicates and event tuples. Our model proposes the use of a semantic measure based on distributional semantics as described in Section 3.1. While typical semantic measures take as input two terms and returns a value in $[0, 1]$, our thematic matcher passes the subscription and event themes as additional parameters along with the terms. The themes are used to adapt the terms meaning vector space before the actual semantic distance is measured as described in Section 4. The various aspects of the model instantiation are discussed in the following sections.

3.1 Distributional Semantics

Distributional semantics is based on the hypothesis that similar and related words appear in similar contexts [15]. Distributional models are quite useful for the task of assessing semantic similarity and relatedness between terms. A semantic measure is a function that quantifies the similarity/relatedness between two terms and typically has its values in the range $[0, 1]$. Distributional models can be constructed automatically from statistical co-occurrence of words in a corpus of documents. The formalism of such a model as a vector space provides a computationally efficient framework for calculating similarity scores.

We scope this paper to the distributional Explicit Semantic Analysis semantic measure *esa* [12] constructed from the Wikipedia corpus as of 2013[2]. However the model is generic and suitable for other measures too. In a nutshell, Wikipedia-based *esa* builds an index of words based on the Wikipedia articles they appear in as shown in Figure 5. 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 or Euclidean distance between the two vectors representing the two terms. In our thematic parametric vector space model, the *esa* measure is parametrized also with the theme tags. They are used to project the terms vectors to get more domain-specific meaning vectors and then are passed to the distance function as illustrated in Figure 5 and detailed in Section 4.

3.2 Themes

We define a theme as a set of terms that describe the content of an event or a subscription. For instance, the set {*'energy', 'appliances', 'building'*} refers to an event which convey energy consumption of appliances in a building. A theme combined with the actual content form an approximation of the meaning of concepts meant to be exchanged in addition to the actual symbols, i.e. words, used to represent attributes and values. A theme is a lightweight method to convey semantics when combined with a semantic model such as distributional semantics. At the same time, themes are meant to be used in situations where little or no agreements can be achieved on a fixed taxonomy.

Event publishers associate their events with a number of terms that describe their payload. Subscribers also associate their subscriptions with a number of terms that clarify their interests. If agreements on themes can be achieved then a theme is decided for each event type. If agreements cannot be assumed then event publishers and subscribers freely add themes that better represent their artifacts.

3.3 Event Model

The event model used in this work is an attribute-value model but the discussion is as relevant to other models such

[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download

as hierarchical or graph-based event models. Each event is a pair of two sets: a set of theme tags and a set of tuples. Each theme tag is a single-word or a multi-word term. Each tuple consists of an attribute-value pair. No two distinct tuples can have the same attribute. An example energy consumption event is represented as follows:

> ({*energy, appliances, building*},
> {**type**: increased energy consumption event,
> **measurement unit**: kilowatt hour,
> **device**: computer, **office**: room 112})

The formal definition of the event model is as follows: let E be the set of all events, let TH be the set of all possible theme tags, and let A and V be the sets of possible attributes and values respectively. Let AV be the set of possible attribute-value pairs, i.e. tuples, such that $AV = \{(a, v) : a \in A \wedge v \in V\}$. An event $e \in E$ is a pair (th, av) such that $th \subseteq TH$ and $av \subseteq AV$ are the set of theme tags and the set of tuples respectively.

3.4 Language Model

Each subscription is a pair of two sets: a set of theme tags and a set of conjunctive attribute-value predicates. Each theme tag is a single-word or a multi-word term. Each predicate uses the equality operator to signify exact equality or approximate equality when indicated. Other Boolean and numeric operators such as $!=$, $>$, and $<$ are kept out of the language for the sake of discourse simplicity. Each predicate consists of an attribute, a value, and specifications of the semantic approximation for the attribute and the value. The most notable feature of the language is the *tilde* $\sim$ operator which helps specify the approximation for an attribute/value when it follows it. An example subscription to energy usage events is as follows:

> ({*power, computers*},
> {**type**= increased energy usage event∼,
> **device**∼= laptop∼, **office**= room 112})

The author of the subscription specifies that the device can be a *'laptop'* or something related semantically to *'laptop'*. The subscription also states that the attribute *'device'* itself can be semantically relaxed. However, it states that the event's *'office'* must be exactly *'room 112'*, etc.

The formal definition of the language model is as follows: let S be the set of subscriptions, let TH be the set of all possible theme tags, and let A and V be the sets of possible attributes and values respectively which can be used in a subscription. Typically there are no restrictions on A or V and the user is free to use any term or combination of terms. Each predicate is a quadruple which consists of the attribute, the value, and whether or not the attribute/value are approximated. Let P be the set of possible predicates, thus $P = \{p : p = (a, v, app_a, app_v) \in A \times V \times \{0, 1\}^2\}$. A subscription $s \in S$ is a pair (th, pr) where $th \subseteq TH$ and $pr \subseteq P$ are the set of theme tags and the set of predicates respectively. The *degree of approximation* is the proportion of relaxed attributes and values. An exact subscription has 0% degree of approximation.

3.5 Matching Model

An approximate semantic single event matcher $\mathcal{M}$ decides on the semantic relevance between a subscription s and an event e based on the semantic mapping between attribute-value predicates of s and attribute-value tuples of e. An example mapping between the event in Section 3.3 and the approximate subscription in Section 3.4 is as follows:

> σ ={(**type=increased energy consumption event** ↔ type:increased energy usage event),
> (**device∼ = laptop∼** ↔ device:computer),
> (**office = room 112** ↔ office: room 112)}

$\mathcal{M}$ works in two modes: the top-1 mode which decides on the most probable mapping between s and e, and the top-k mode which decides on the top-k probable mappings to be used later for complex event processing. It has been shown in [13] that producing the top-k mappings increases the chance of hitting the correct mapping.

The formal definition of matching is as follows: let $C = s \times e$ be the set of all possible correspondences between predicates of s and tuples of e. $\forall c = (p, t) \in C \Rightarrow p \in s \wedge t \in e$. $\Sigma = 2^C$ is the power set of C and represents all the possible mappings between s and e. There are exactly n correspondences in any valid mapping σ where n is the number of predicates in the subscription s.

For any valid mapping σ a probability function quantifies the probability of every predicate-tuple correspondence $(p, t) \in \sigma$ such as (**device = laptop∼** ↔ device: computer). There also exists a probability function which quantifies the probability of the overall mapping σ among other possible mappings. Both functions form probability spaces $\mathcal{P}_\sigma$ and $\mathcal{P}$. In this paper, all probabilities are calculated based on the combined similarity matrix which is based on the thematic pairwise attributes or values semantic relatedness scores. Thematic semantic relatedness measure is discussed in Section 4. For more details on the generic matcher model and detailed evaluation of top-1 and top-k modes, please refer to [16].

4. PARAMETRIC VECTOR SPACE MODEL

We introduce the concept of a Parametric Vector Space Model (PVSM). Vector space models are widely used in information retrieval and known to be computationally efficient. Thus, we propose an extension suitable for event processing where time efficiency is a requirement. Figure 5 shows the main elements of the parametric space. Building the PVSM is identical to building the non-thematic distributional space model based on indexing the corpus. Nonetheless, vectors in PVSM are projected into thematic dimensions passed as parameters before being used as discussed in the following subsections.

4.1 Distributional Vector Space Model

Given a set of documents D, each document is tokenized into terms, stop words are removed, and an inverted index is built to have an entry for each term [6], step 1 in Figure 5. The inverted index encodes a vector space model whose basis is the set of unit vectors that represent the documents, i.e. $\{\vec{d_i} : d_i \in D\}$. Each term t is then represented as a weighted vector $\vec{v_t}$ in the vector space as shown in Equation 1.

$$\vec{v_t} = \sum_{i=1}^{i=|D|} w_{ti}\vec{d_i} \tag{1}$$

We use the Term Frequency Inverse Document Frequency (TF/IDF) weighting scheme which gives more weight to a

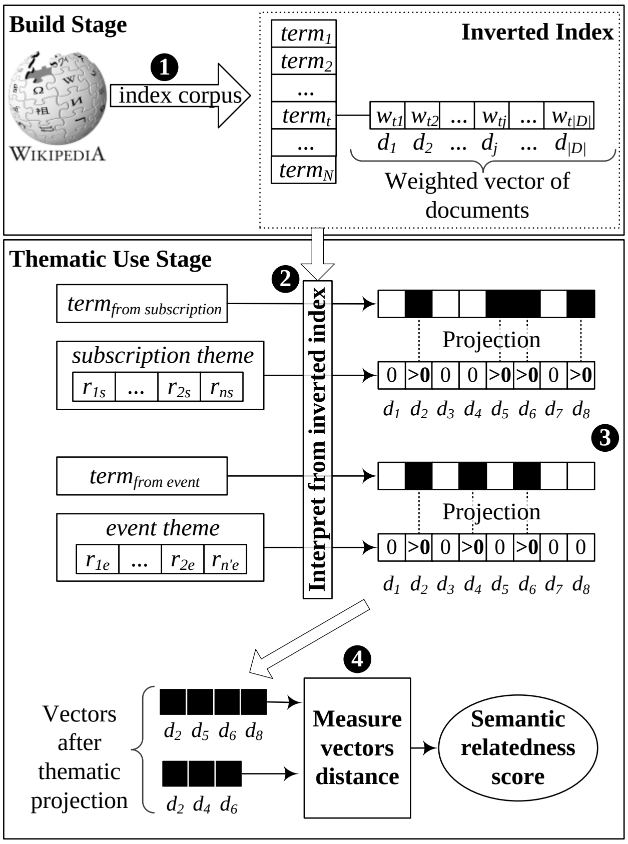

Figure 5: Parametric distributional vector space.

term if it appears more often in a document and less often in other documents. It is important to keep the raw tf and idf values for each pair (term, document) in the inverted index so they can be used later for thematic projection. TF/IDF scheme is shown in Equations 2, 3, and 4.

$$tf(t,d) = 0.5 + \frac{0.5 \times freq(t,d)}{max\{freq(t',d) : t' \in d\}} \tag{2}$$

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{3}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{4}$$

4.2 Thematic Projection

At the usage stage, the ultimate goal is to measure the semantic relatedness between two terms $term_s$ and $term_e$ given the subscription and the event themes th_s and th_e respectively. Given a term and a theme, the key operation is to use the theme to filter the space into a thematic subspace. The thematic space basis is the set of documents that define the theme representative tags.

The thematic basis can be found by getting the vector representation of the theme, step 2 in Figure 5, and then the documents where its weights are greater than 0, step 3 in Figure 5. Given the new basis, the term vector is transformed to have 0 components for documents not in the thematic basis and to have new $tfidf$ weights for the basis documents as the overall number of documents is now different from $|D|$. These steps are shown in Algorithm 1. Projection can be computed in $O(|D|)$ time if all vectors components are stored and $O(|V|)$ where V is the non-zero components if only those are stored in the index.

ALGORITHM 1: Thematic projection

```
Input: a term $t$, a set of theme tags $th$, parametric
       distributional vector space $PVSM$
Result: thematic projection vector $\vec{t_{th}}$ of $t$ given $th$
begin
    $\vec{t} \longleftarrow$ distributional vector of $t$ from $PVSM$;
    $\vec{th} \longleftarrow$ distributional vector of $th$ from $PVSM$;
    for $d \in D$ $s.t.$ $\vec{th}_d = 0$ do
        $\vec{t_{th}}_d \longleftarrow 0$;
    end
    for $d \in D$ $s.t.$ $\vec{th}_d > 0$ do
        $tf \longleftarrow$ original $tf(t,d)$ from $PVSM$;
        $idf \longleftarrow log\frac{|\{d \in D : \vec{th}_d > 0\}|}{|\{d \in D : \vec{th}_d > 0 \wedge \vec{t}_d > 0\}|}$; /* recalculate idf */
        $\vec{t_{th}}_d \longleftarrow tf \times idf$;            /* update weight */
    end
    return $\vec{t_{th}}$;
end
```

4.3 Distance and Semantic Relatedness

Let T be the set of terms, and TH the set of all possible thematic tags. We define the semantic measure sm as a function that operates on a pair of terms associated with their themes such that $sm : T \times 2^{TH} \times T \times 2^{TH} \rightarrow [0,1]$. Given two terms t_s and t_e from a subscription and an event respectively and their associated themes $th_s \in 2^{TH}$ and $th_e \in 2^{TH}$ respectively, sm works by finding the thematic projections $\vec{t_{s_{th_s}}}$ and $\vec{t_{e_{th_e}}}$ and then calculating the vector distance between the resulting projected vectors, step 4 in Figure 5.

We use the *Euclidean* distance to measure projected vectors distance as defined in Equation 5.

$$dis(\vec{a},\vec{b}) = \sqrt{\sum_{i=1}^{i=|D|} (\vec{a}_i - \vec{b}_i)^2} \tag{5}$$

Semantic relatedness is estimated to be the opposite of the distance and can be calculated as defined in Equation 6.

$$relatedness(\vec{a},\vec{b}) = \frac{1}{dis(\vec{a},\vec{b}) + 1} \tag{6}$$

The more filtering that occurs during thematic projection due to smaller themes, the less time is required for computing relatedness.

5. EVALUATION

To evaluate the thematic approach, we compare it with non-thematic approximate semantic event processing. Evaluation is concerned with two metrics: matching quality and time efficiency. In previous work [16] we compared the non-thematic approximate approach with a concept-based approach that uses query rewriting using WordNet [20]. Experiments were conducted with 10 sets of $10 - 100$ approximate subscriptions of 50% degree of approximation with *esa*. Results show that the approximate matching model delivers $94\% - 97\%$ matching quality, higher than the $89\% - 92\%$ delivered by the WordNet rewriting approach.

Table 2: Base Concepts to Evaluate Effectiveness

	Ground Truth Relevant Events	**Ground Truth** Irrelevant Events
Matcher Relevant Events	TP (True Positive)	FP (False Positive)
Matcher Irrelevant Events	FN (False Negative)	TN (True Negative)

The rewriting approach outperforms the approximate approach in throughput when the pair-wise semantic relatedness scores are calculated at run-time. However, the approximate model based on precomputed *esa* scores outperforms in throughput with around $91,000$ events/sec compared to around $19,100$ events/sec on average. Distributional semantics-based approximation is based on a very loose model of semantic coupling which scales to heterogeneous environments. That is not the case for rewriting with knowledge bases as building them is time consuming and establishing agreements is granular and difficult to achieve.

In this paper we generate a large event set with a particular theme as well as a set of subscriptions which assume no semantic agreements and 100% degree of approximation. We compare the thematic matcher with the non-thematic matcher when different theme tags are used. Evaluation metrics and detailed methodology are described in the following subsections.

5.1 Evaluation Metrics

Evaluation metrics can be classified into two categories: effectiveness and efficiency metrics [4]. Effectiveness metrics measure the quality of event matching. That requires a ground truth which divides events into relevant and irrelevant with respect to each subscription. Table 2 shows the base concepts needed for evaluating effectiveness. For those to exist, the resulting events from the matcher must be divisible into two distinct sets of relevant and irrelevant events. In the case of the approximate matcher which assigns probabilities to events with respect to a subscription, the two sets can be achieved by ranking and cutting off using recall levels. *Precision*, *Recall*, and the combined F_1Score have been used for effectiveness evaluation.

Precision measures the proportion of relevant events discovered by the matcher with respect to all the discovered events such that $Precision = TP/(TP + FP)$. *Recall* measures the proportion of relevant events discovered by the matcher with respect to all the known relevant events from the ground truth such that $Recall = TP/(TP + FN)$. Precision and recall are calculated for the whole set of subscriptions by averaging the precision and recall achieved for all individual subscriptions respectively. The F_1Score equally combines *Precision* and *Recall* such that $F_1Score = (2 \times Precision \times Recall)/(Precision + Recall)$. F_1Score is computed at 11 recall points, $\{0, 0.1, 0.2, ..., 1.0\}$, to cover all the precision-recall curve without using thresholds and the maximal F_1Score is then used. The metric used for evaluating time efficiency is *Throughput* defined as $Throughput = (Number\ of\ processed\ events)/(Time\ unit)$.

5.2 Methodology

The evaluation methodology for effectiveness outlined in Figure 6 is based on schema matching/mapping methodologies [4] where the task is to find the best mapping between a source schema and a target schema. Within the context of event matching we use approximate subscriptions and events. Specifying the ground truth mappings is a challenge for large sets of events and subscriptions.

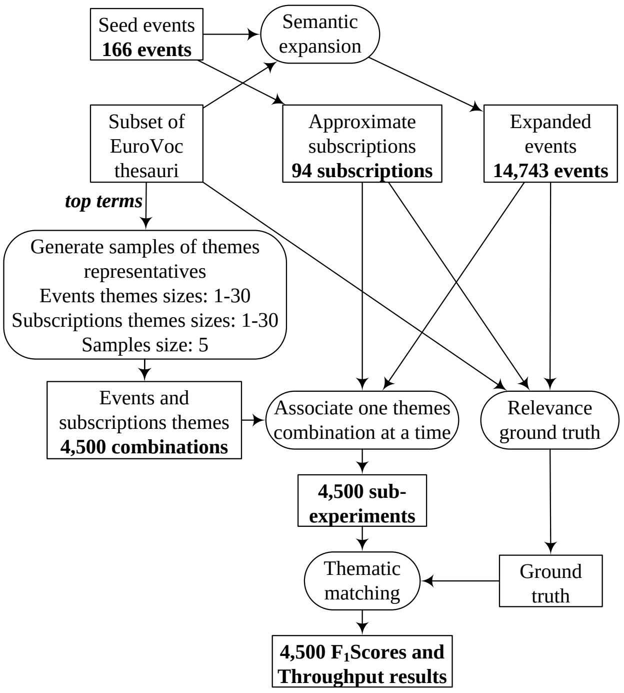
Seed events
166 events
Semantic expansion
Subset of EuroVoc thesauri
Approximate subscriptions
94 subscriptions
Expanded events
14,743 events
top terms
Generate samples of themes representatives
Events themes sizes: 1-30
Subscriptions themes sizes: 1-30
Samples size: 5
Events and subscriptions themes
4,500 combinations
Associate one themes combination at a time
Relevance ground truth
4,500 sub-experiments
Thematic matching
Ground truth
4,500 F1Scores and Throughput results

Figure 6: Evaluation methodology.

In recent years there has been a trend towards synthetic evaluation [4]. Similarly to the idea in STBenchmark [1], we start with pairs of exact subscriptions and events with a known ground truth which is simply the result of exact matching. We then apply a semantic expansion transformation to the events and the subscriptions based on a thesaurus, similarly to the synonyms transformation in eTuner [18]. The ground truth is updated accordingly. We use the *EuroVoc*[3] thesaurus for themes and ground truth generation as it has many domains and can be used for semantic expansion according to specific themes. EuroVoc is a multilingual and multidisciplinary thesaurus that provides common lexis to cover the activities of the European Union.

5.2.1 Generation of the Seed Event Set

To create a heterogeneous IoT environment, we have created a dataset of events using a set of real-world datasets. Seed events have been synthesized from a set of IoT sensors identical to those deployed in the SmartSantander smart city project [23] and the Linked Energy Intelligence (LEI) dataspace [8]. SmartSantander proposes a city-scale experimental research testbed for IoT applications and services based on sensors deployed in a set of European cities. The LEI project targets smart buildings for energy saving purposes. The used sensor capabilities are shown in Table 3.

A set of car brands from the Yahoo! directory[4] is used

[3] ©European Union, 2014, http://eurovoc.europa.eu/
[4] http://dir.yahoo.com/recreation/automotive/makes_and_models

Table 3: Sensor Capabilities

Sensor Capabilities
solar radiation, particles, speed, wind direction, wind speed, temperature, water flow, atmospheric pressure, noise, ozone, rainfall, parking, radiation par, co, ground temperature, light, no_2, soil moisture tension, relative humidity, energy consumption, cpu usage, memory usage

to generate vehicle mobile sensors platforms. A set of appliances from the BLUED KDD dataset are used as indoor platforms [2]. For indoor locations, rooms from the DERI Building[5] have been used. For geographical locations the SmartSantander project locations as well as Galway City have been used. The seed event generation is done by randomly combining various attributes and values from the aforementioned datasets. A set of 166 seed events has been used to generate events for the experiments. An example seed event generated is as follows:

{**type**: increased energy consumption event,
measurement unit: kilowatt hour, **device**: laptop,
desk: desk 112c, **room**: room 112, **zone**: building,
city: Galway, **country**: Ireland, **continent**: Europe}

5.2.2 *Semantic Expansion of Seed Events*

The purpose of semantic expansion of seed events is to generate a large amount of events for evaluation where semantic heterogeneity holds. The EuroVoc thesaurus has been used and specifically its micro-thesauri belonging to domains *'transport'*, *'environment'*, *'energy'*, *'geography'*, *'education and communications'*, and *'social questions'*. This is because those micro-thesauri conform to the theme of the events used in the experiments. A set of $14,743$ expanded events of a length up to 10 tuples has been generated starting from seed events by replacing one or more terms in an event's tuples by synonyms or related terms from the thesaurus. An example event resulting from semantically expanding the seed event in Section 5.2.1 is as follows:

{**type**: increased energy consumption event,
measurement unit: kilowatt hour, **device**: laptop,
desk: desk 112c, **room**: room 112, **zone**: building,
urban area: Galway, **country**: Eire,
continent: European countries}

5.2.3 *Generation of Approximate Subscription Set and Ground Truth*

A set of 94 exact subscriptions are generated by randomly picking a number of tuples from the seed events and turning them into exact subscriptions. A set of 94 approximate subscriptions are then generated by introducing the *tilde* $\sim$ operator into all the predicates in the exact subscriptions to exclude the non-approximation effect on the results. The approximate subscriptions are equivalent to about $48,000$ subscriptions which would be needed by a non-approximate approach to cover events heterogeneity. An example approximate subscription resulting from relaxing all predicates of the exact subscription is as follows:

{**type**$\sim$: increased energy consumption event$\sim$,
device$\sim$: laptop$\sim$, **floor**$\sim$: ground floor$\sim$}

The resulting relevance function between approximate subscriptions and expanded events is isomorphic to a basic exact ground truth function between exact subscriptions and seed events, thus it is an exact relevance function. As a result, an expanded event is relevant to an approximate subscription if it exactly matches the subscription or a version of it which results from it by replacing the approximated parts with related terms from the thesaurus used for semantic expansion.

5.2.4 *Generation of Theme Tags*

The target of this step is to associate events and subscriptions with themes tags. EuroVoc has *top terms* for each of its micro-thesauri. We randomly pick from the top terms associated with the domains *'transport'*, *'environment'*, *'energy'*, *'geography'*, *'education and communications'*, and *'social questions'* which are originally used to expand the event set. For each sub-experiment two sets of representative tags are chosen to represent the subscriptions theme and the events theme. The purpose is to study the behavior of the thematic approximate matcher with different combinations of themes tags. An example subscription theme tags set from EuroVoc of size 2 is {*land transport*, *protection of nature*}.

Given the events and subscriptions sets, various combinations of theme tags have been associated to them. For each combination, we have a sub-experiment which gives an F_1Score and a throughput result. In every combination, the event theme tags set contains the subscription theme tags set or vice versa. Each combination is defined by the size of the event and the subscription themes. For example, a $3-2$ combination means that the event theme contains 3 terms while the subscription theme contains 2 terms and the former contains the latter.

For each combination of sizes, we have a random sample of 5 different pairs of theme tags sets. The experiment has been conducted with different sizes of 1 to 30 tags for subscriptions and 1 to 30 tags for events. This gives $30 \times 30 \times 5 = 4,500$ sub-experiments. The thematic matcher was executed in each sub-experiment to give F_1Score and throughput results. The choice of the sample size is due to the high dimensionality of the experiments which poses practical constraints. We think that future work shall use more resources to allow experimentation with larger samples.

5.2.5 *Baseline*

Given the generated events and subscriptions sets, a non-thematic approximate matcher with domain-independent *esa* has been used [16]. The matcher gives 62% of F_1Score and a throughput of 202 events/sec averaged over 5 runs.

5.3 Results

The following subsections discuss the effectiveness and efficiency results. All experiments have been conducted on a Windows 7 machine, with an Intel Core i7-3520 2.90 GHz CPU and 8GB of RAM running JVM 1.7.

5.3.1 *Effectiveness*

Each cell in Figure 7 represents the average F_1Score of the sample of 5 sub-experiments, each of which uses a different combination of events and subscriptions themes tags. For instance, the sub-experiments of the cell in the 2^{nd} column

[5] http://lab.linkeddata.deri.ie/2010/deri-rooms

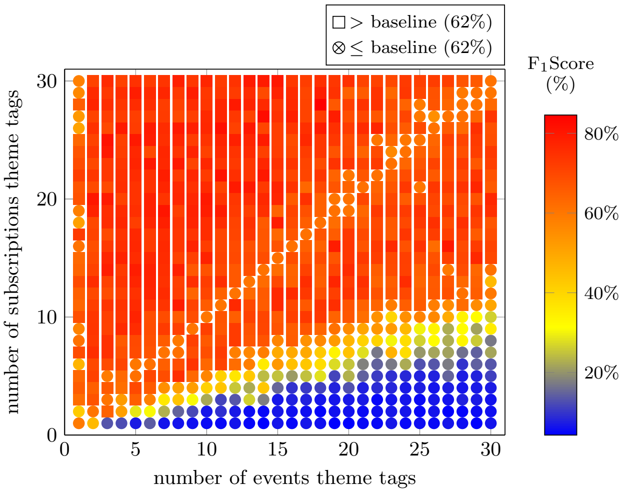

Figure 7: Effectiveness of thematic matcher.

and 10^{th} row from the bottom left, all use 2 terms to describe events theme and 10 terms to describe subscriptions theme and the event theme terms set is a subset of the subscription theme terms set. The sub-experiments of the cell in the 10^{th} column and 10^{th} row from the bottom left, all use 10 terms to describe events theme and 10 terms to describe subscriptions theme and the event theme terms are the same as the subscription theme terms. Square cells are sub-experiments which exceed the baseline while circular ones score below the baseline. Cell color reflects the average F_1Score for the sample of combinations for that cell. Colors range from blue (low F_1Score) to red (high F_1Score).

Figure 7 shows that thematic matching outperforms non-thematic matching in F_1Score for more than 70% of combinations with scores $62\%-85\%$ and an average of 71% versus 62% for the baseline. Those are more concentrated in the upper left two thirds of Figure 7. F_1Score on the diagonal line is also a little less for the thematic matcher, $59\%-62\%$ versus 62%, suggesting that the projection stage of the vector space by same tags seems to be less discriminative as opposed to using different tags which could disambiguate attributes/values better. Thematic matching performs worse when the number of thematic tags is very small, e.g. using just one term as a theme tag. Also, in the bottom triangular half of the figure with F_1Score widely ranging from 4% to 62%. Larger themes for subscriptions quickly improve effectiveness as opposed to an opposite effect by event themes. That reflects the asymmetric relationship between the many heterogeneous events versus fewer subscriptions. Thus more terms are needed in subscription themes to discriminate relevant events.

Figure 8 illustrates the standard deviation (standard error) of the samples conforming to each set of 5 combinations. The average standard error is 7% of F_1Score in effectiveness. Most of this error is around sub-experiments of medium F_1Scores where it reaches values around $10\%-25\%$. Very small errors are concentrated around the sub-experiments of very low F_1Scores but those are not of concern as theme combinations conforming to such areas of Figure 7 should be avoided. More importantly, error converges to smaller values around 7% for sub-experiments of high F_1Scores which mainly exceed the baseline. This suggests that the experiments are more predictive for higher F_1Scores and the areas of Figure 7 which outperforms the non-thematic approach are more probable to outperform it also in other samples.

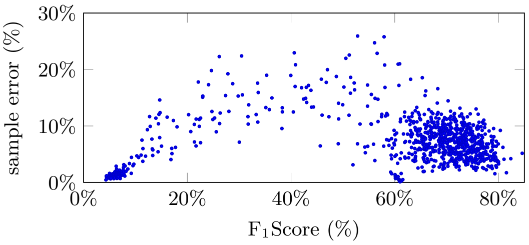

Figure 8: Effectiveness sample error.

5.3.2 *Time Efficiency*

Figure 9 shows the average throughput for each combination of events and subscriptions theme tags. It suggests that the thematic approach outperforms the non-thematic matcher for more than 92% of the sub-experiments, with throughput of $202-838$ and an average of 320 versus 202 events/sec. Better throughput is due to the thematic filtering of the space during the thematic projection phase which saves time during semantic relatedness calculation. That has less effect given more tags towards the top right corner with throughput as low as 95 events/sec.

Figure 9 shows that throughput decreases gradually when larger sets of theme tags are used to describe events and subscriptions due to less thematic filtering. The last half of the diagonal line shows a drop in throughput, $95-177$ versus 202 events/sec, as two equal sets of thematic tags for events and subscriptions causes more common dimensions for the semantic measure to be calculated and thus more time is needed for calculation.

Figure 10 shows that few sub-experiments outliers (around 5%) have high standard deviation ranging from 20 to 240 events/sec. The outliers can be explained by rare terms that do not exist in the original indexed corpus which causes the space to be filtered completely and results in a very different time consumption behavior from other combinations in the same sample. This causes higher errors and less predictability. However, most other sub-experiments have a standard error around the average of 10 events/sec which is small compared to the overall throughput. Most of small errors are identified around sub-experiments with throughput from $200-600$ events/sec which is mainly above the non-thematic baseline. This shows that throughput results are well predictive and should be expected in other samples of subscriptions and events theme combinations.

In previous work, we discussed less degrees of approximations when some agreements can be assumed and throughput of a magnitude of thousands events/sec was achieved [16]. Experiments here represent a worst case scenario with 100% degree of approximation and were conducted on a single laptop. We think there are further opportunities to optimize the matcher with commonalities, evaluation order, caching, and indexing techniques to improve efficiency.

5.3.3 *Discussion*

Results show that the thematic approach is limited when users can provide only a small number of tags for subscrip-

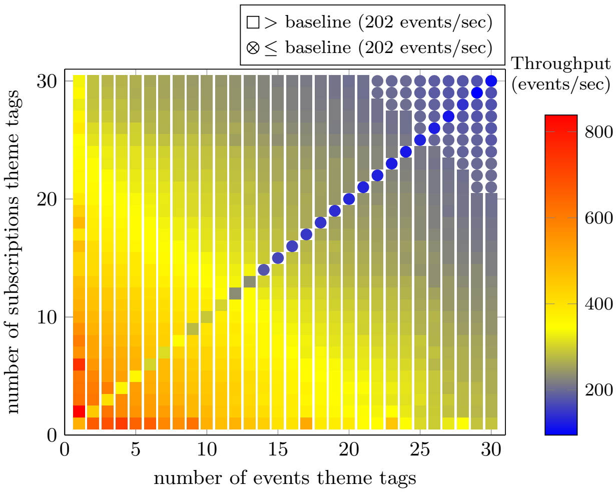

Figure 9: Throughput of thematic matcher.

tions, 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 and throughput together with less error rates. That is concentrated in the middle to upper left side of Figures 7 and 9. We think that this is a lightweight amount of terms that events and subscriptions authors can associate with their artifacts.

For containment between subscriptions themes and events themes to hold, it can be handled in two ways:

- Event sources and consumers loosely agree on terms to use which guarantee containment but causes some semantic coupling.

- Event sources and consumers use more theme tags when no agreement can be achieved in vastly open and decoupled scenarios. Containment and overlap can be assumed to hold due to the distribution of term usage by humans where some terms are more probable to be used by both parties.

6. RELATED WORK

Related work can be recognized in distributed event-based systems and information retrieval communities.

6.1 Semantic Event Processing

A-TOPSS [19] defines an approximate matching model based on fuzzy functions that specify the degree of membership between an event's value and a subscription's filter but without supporting schema approximation. S-TOPSS [22] tackles schema and value semantic matching via agreed-upon ontologies and a system architecture that generates events other than the original ones by replacing concepts with taxonomic concepts. S-TOPSS provides a generic architecture but no concrete model or empirical validation has been discussed. Besides, replicating events with new concepts has the downside of overwhelming the system with a large amount of events. FOMatch [29] proposes the use of fuzzy agreed-upon ontologies. However, it does not free the user from using pre-defined vocabularies.

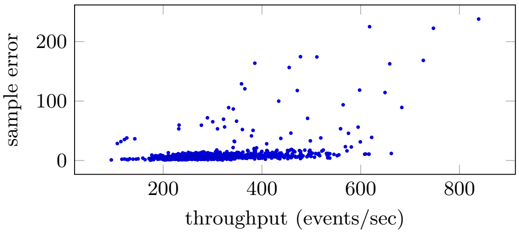

Figure 10: Throughput sample error.

6.2 Uncertainty in Event Processing

A taxonomy for uncertain event processing has been proposed in [25]. It suggests two dimensions for uncertainty: element and origin uncertainty. *Origin uncertainty* deals with the source of uncertainty which may originate from the event *source* or from event *inference*. Our model suggests matching as another type of origin where uncertainty reflects the loose semantic coupling between sources and consumers. A model for complex event processing over uncertain events is proposed in [26]. Single event matching in our model can feed into a complex event processing module.

6.3 Folksonomies

Folksonomies are taxonomies generated by people (folks) in a bottom-up manner. It has been widely adopted as an alternative to top-down agreed upon taxonomies which require effort to agree upon, to maintain, and to use [14]. Gupta et al. recognize user motivations for tagging in [14] such as contribution, sharing, and technological ease. A classification of tags is presented to include content-based, context-based, ownership, organizational, and personal tags among others. Several successful applications of folksonomies are discussed including: search [27], indexing, classification, enhanced browsing, and others. Folskonomies are typically used to index non-structured web pages and images while our work introduces the parametric vector space model which uses tags to extract approximate meanings of structured events data and subscriptions.

7. CONCLUSIONS AND FUTURE WORK

This paper proposes a thematic event processing approach to deal with semantic boundaries arising in large-scale, heterogeneous, and open environments. The proposed approach suggests associating events and subscriptions with tags to describe their semantic themes. The themes represent a lightweight way to communicate event semantics across systems boundaries without the need for granular semantically coupling agreements that limit scalability. Experiments show that the thematic approach outperforms a non-thematic approximate event processing approach in matching effectiveness and throughput for many combinations of theme tags. Future work aims at the study of realistic tagging behavior of users, building an efficient indexing for thematic projection, throughput optimization, and more quantitative aspects of evaluation such as cold start and real-time behavior.

8. ACKNOWLEDGMENTS

This work has been supported in part by the SENSE project and conducted in the Insight Centre for Data Analytics at National University of Ireland, Galway.

9. REFERENCES

[1] B. Alexe, W.-C. Tan, and Y. Velegrakis. STBenchmark: towards a benchmark for mapping systems. *Proceedings of the VLDB Endowment*, 1(1):230–244, 2008.

[2] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges. BLUED: a fully labeled public dataset for event-based non-intrusive load monitoring research. In *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*, 2012.

[3] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787–2805, 2010.

[4] Z. Bellahsene, A. Bonifati, F. Duchateau, and Y. Velegrakis. On evaluating schema matching and mapping. In *Schema Matching and Mapping*, pages 253–291. Springer, 2011.

[5] P. R. Carlile. Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization Science*, 15(5):555–568, 2004.

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

[7] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf. Achieving scalability and expressiveness in an internet-scale event notification service. In *Proceedings of the 19th Annual ACM Symposium on Principles of Distributed Computing*, pages 219–227. ACM, 2000.

[8] E. Curry, S. Hasan, and S. O'Riain. Enterprise energy management using a linked dataspace for energy intelligence. In *Sustainable Internet and ICT for Sustainability (SustainIT)*, pages 1–6. IEEE, 2012.

[9] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec. The many faces of publish/subscribe. *ACM Computing Surveys*, 35(2):114–131, 2003.

[10] P. T. Eugster, R. Guerraoui, and C. H. Damm. On objects and events. In *ACM SIGPLAN Notices*, volume 36, pages 254–269. ACM, 2001.

[11] A. Freitas, J. G. Oliveira, S. O'riain, J. C. Da Silva, and E. Curry. Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data & Knowledge Engineering*, 88:126–141, 2013.

[12] 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*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007.

[13] A. Gal. Managing uncertainty in schema matching with top-k schema mappings. In *Journal on Data Semantics VI*, pages 90–114. Springer, 2006.

[14] M. Gupta, R. Li, Z. Yin, and J. Han. Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, 12(1):58–72, 2010.

[15] Z. S. Harris. Distributional structure. *Word*, 10:146–162, 1954.

[16] S. Hasan and E. Curry. Approximate semantic matching of events for the internet of things. *ACM Trans. Internet Technol.*, 14(1):2:1–2:23, Aug. 2014.

[17] 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*, DEBS '12, pages 252–263, 2012.

[18] Y. Lee, M. Sayyadian, A. Doan, and A. S. Rosenthal. eTuner: tuning schema matching software using synthetic scenarios. *The VLDB Journal-The International Journal on Very Large Data Bases*, 16(1):97–122, 2007.

[19] H. Liu and H.-A. Jacobsen. A-TOPSS: a publish/subscribe system supporting approximate matching. In *Proceedings of the 28th international conference on Very Large Data Bases*, VLDB '02, pages 1107–1110. VLDB Endowment, 2002.

[20] G. A. Miller. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[21] OECD. Machine-to-Machine Communications: Connecting Billions of Devices. http://www.oecd-ilibrary.org/. 2012.

[22] 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*, VLDB '03, pages 1101–1104. VLDB Endowment, 2003.

[23] 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*, pages 1–8. IEEE, 2011.

[24] C. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1949.

[25] S. Wasserkrug, A. Gal, and O. Etzion. A taxonomy and representation of sources of uncertainty in active systems. In O. Etzion, T. Kuflik, and A. Motro, editors, *NGITS*, volume 4032 of *Lecture Notes in Computer Science*, pages 174–185. Springer, 2006.

[26] S. Wasserkrug, A. Gal, O. Etzion, and Y. Turchin. Complex event processing over uncertain data. In *Proceedings of the Second International Conference on Distributed Event-based Systems*, DEBS '08, pages 253–264, New York, NY, USA, 2008. ACM.

[27] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, pages 155–162. ACM, 2008.

[28] L. Zeng and H. Lei. A semantic publish/subscribe system. In *IEEE International Conference on E-Commerce Technology for Dynamic E-Business*, pages 32–39, 2004.

[29] W. Zhang, J. Ma, and D. Ye. FOMatch: A fuzzy ontology-based semantic matching algorithm of publish/subscribe systems. In *The International Conference on Computational Intelligence for Modelling Control Automation*, pages 111–117, 2008.