Approximate Semantic Matching of Heterogeneous Events

Souleiman Hasan
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
souleiman.hasan@deri.org

Sean O'Riain
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
sean.oriain@deri.org

Edward Curry
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
ed.curry@deri.org

ABSTRACT

Event-based systems have loose coupling within space, time and synchronization, providing a scalable infrastructure for information exchange and distributed workflows. However, event-based systems are tightly coupled, via event subscriptions and patterns, to the semantics of the underlying event schema and values. The high degree of semantic heterogeneity of events in large and open deployments such as smart cities and the sensor web makes it difficult to develop and maintain event-based systems. In order to address semantic coupling within event-based systems, we propose vocabulary free subscriptions together with the use of approximate semantic matching of events. This paper examines the requirement of event semantic decoupling and discusses approximate semantic event matching and the consequences it implies for event processing systems. We introduce a semantic event matcher and evaluate the suitability of an approximate hybrid matcher based on both thesauri-based and distributional semantics-based similarity and relatedness measures. The matcher is evaluated over a structured representation of Wikipedia and Freebase events. Initial evaluations show that the approach matches events with a maximal combined precision-recall F_1 score of 75.89% on average in all experiments with a subscription set of 7 subscriptions. The evaluation shows how a hybrid approach to semantic event matching outperforms a single similarity measure approach.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability---*data mapping, interface definition languages*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval---*information filtering*.

General Terms

Algorithms, Experimentation, Human Factors, Languages.

Keywords

Approximate Event Matching, Semantic Decoupling, Semantic Event Matching.

DEBS'12, July 16–20, 2012, Berlin, Germany.
Copyright 2012 ACM 978-1-4503-1315-5...$10.00.

1. INTRODUCTION

Event-based technology is becoming more widely needed with the rise of new applications ranging from smart homes to smart cities and the Internet-of-Things [1]. Event-based systems enable a decoupled mode of interaction between participants making it suitable for large scale distributed environments [10]. There are estimates that by the end of 2020 fifty billion devices will be connected to mobile networks [22] which would push event-based technology to its limits.

While event-based systems are decoupled in space, time, and synchronization [10], scaling out to include participants from diverse domains poses a challenge with the semantic interpretation of events. Current systems assume mutual agreement on event semantics which adds explicit dependencies between interacting parties. This ties event subscriptions and processing languages to crisp and well understood schema and semantics of events. This can limit the scalability of an event-based system to that of the events for which the schema and semantic interpretation is known. The requirement of an upfront understanding of the event semantics creates semantic coupling that can limit scalability especially in environments with high levels of semantic heterogeneity. It also puts a barrier between non-technical users who do not fully understand the used semantics and event-based systems. That constrains usability by non-technical users and limits it to IT specialists. Thus, there is a need to recognize event semantics as a fourth dimension of coupling if event-based systems are to scale out to highly heterogeneous environments such as the Internet of Things [1].

Semantic decoupling of events and user's subscriptions requires an appropriate method for matching and processing of events. One approach to event matching is approximate semantic matching which uses a mechanism for ranking events according to their relevance to users' subscriptions. We propose in this paper a model for approximate semantic matching that addresses event semantic decoupling requirement. We instantiate our model using a hybrid matching approach based on both thesauri and distributional semantics-based semantic similarity and relatedness measures. A novel evaluation that leverages heterogeneous real world events created by human and extracted from Wikipedia and Freebase is conducted with promising matching results.

The rest of this paper is organized as follows: Section 2 motivates the problem of semantic coupling in an enterprise scenario and an open web scenario while Section 3 discusses decoupling in event-based systems. Section 4 explains the proposed approach and Section 5 details an instantiation of the proposed event, subscription, and matching models. The approach is evaluated in Section 6. Section 7 analyses related work. Potential future directions are identified in Section 8, and Section 9 concludes the paper.

2. MOTIVATIONAL SCENARIOS

2.1 Enterprise Scenario

The chief sustainability officer (CSO) is a part of the upper management and responsible for the company social responsibility programmes. The CSO is interested in a simple metric that gives in real-time the company's performance from a carbon emissions perspective with regard to international standards. The CSO is not a technical person so the task is forwarded to the IT department which starts identifying the different potential sources that affects the companies CO_2 [8].

A medium size organization typically has multiple information systems to manage assets, human resources, orders, etc. Heating, ventilation, and air conditioning (HVAC) are managed by a building management system. Energy consumption sensors exist for lights, laptops and data centre. The IT department instruments different emitters with sensors that publish events to an event-based infrastructure. Because energy consumption information comes from heterogeneous sources and generated by devices from different manufactures, it is highly likely that different schemas and values are used. They might use the terms *"energy consumption"* and *"energy usage"* to refer to the same thing. Locations of devices might be described differently as *"rooms"*, *"spaces"*, *"wings"*, etc. A web service from the power utility is used to determine the carbon emissions from power usage. The IT department also creates a rule-based situation assessment (SA) agent to consume raw events, aggregate events according to the different schemas and values and generate overall performance events which are consumed by a dashboard that is shown to the CSO.

The diversity of schema and values results in a large number of rules to process events. That makes the cost of maintainability of the event infrastructure very high when changes in event schemas or value semantics occur or if a new event source is added or changed. E.g. if the external web service starts using *"wind"* instead of *"renewable"*, the SA agent will not be able to match the web service events. The SA agent might stop working for a while until the IT specialists determine the reason and make the necessary changes. Similarly, when a new set of smart fridges is added in the building and they start publishing events with the term *"kitchen"* instead of *"room"*, they will not be accounted automatically in the SA node until special rules are manually added for them.

2.2 Open Web Scenario

A tourist agency is running a website that gathers real-time feeds from the web about interesting events such as sporting games, concerts, circuses, etc. The site allows users to register their interests in some types of events with some characteristics in their planned destinations of trips. The website subscribes to RSS web feeds from thousands of sources such as museums websites, football clubs websites and others. Feeds contain typical RSS items such as *"title"* and the publication date (*"pubDate"*). They may also contain other information items like *"namespace1#club"* or *"namespace2#team"* that conform to the publishers' own descriptions of football matches. When a user subscribes to the agency's website, she prefers using expressions with no restrictions on possible vocabulary that can be used, such as the subscription in Example 1.

Example 1

event type "Football Match"

event team "Barcelona"

Since feeds use different terms such as *"Soccer Match"* instead of *"Football Match"*, *"club"* instead of *"team"* or *"FCB"* instead of *"Barcelona"*, the user misses some events that are relevant to the subscription if the website assumes conjunction between the statements. If the website assumes disjunction, the user may get many events that are played by some team from *Barcelona* but are not football matches. They would be considered equally relevant by the website although the user may want to have basketball games played by *Barcelona* ordered first if no football matches are detected.

3. SEMANTIC COUPLING WITHIN EVENT SYSTEMS

The event-based interaction paradigm is based on decoupling producers and consumers of events. The main advantage of decoupling the production and consumption of events is an increased scalability by "removing explicit dependencies between the interacting participants" [10]. The three common dimensions of coupling between event producers and consumers are space, time and synchronization:

- **Space decoupling** suggests that the interacting parties do not need to know each other. Publishers do not hold references to consumers or know how many of them are actually interacting and vice versa.
- **Time decoupling** means that participants do not need to be actively involved in the interaction at the same time.
- **Synchronization decoupling** suggests that event producers are not blocked while producing events and consumers get notified of an event occurrence while performing some concurrent activity [10].

However, event-based systems that support space, time and synchronization dimensions of decoupling can be still tightly coupled by the semantic of events they exchange. If an event system assumes mutual agreement on event types, properties, and values, this agreement is an explicit dependency between parties. Semantic coupling limits the scalability of event-based systems within deployment environments with high-levels of event heterogeneity. In these environments there is a large cost to define and maintain the whole subscriptions and rules needed by event consumers.

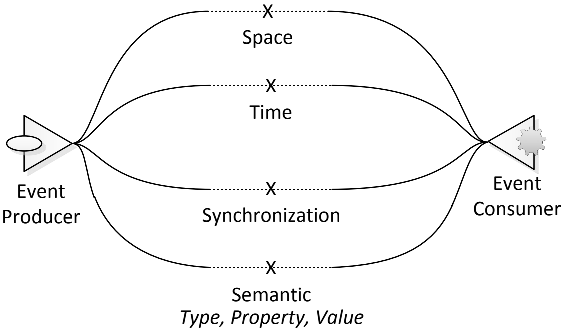

Figure 1. Four dimensions of event decoupling

Mutual agreement between event producers and event consumers suggests a semantic coupling that has three dimensions:

- **Event type semantic coupling** occurs when participants agree on the name of a class of event instances.
- **Event properties semantic coupling** occurs when participants agree on naming of specific attributes and characteristics of an event class.
- **Event values semantic coupling** occurs when participants agree on the interpretation and the set of values that an event instance can have for a property.

Thus, the core requirement tackled in this paper is event semantic decoupling. It aims at adding a fourth dimension of decoupling in order to enforce the principle of removing explicit dependencies between event producers and consumers as illustrated in Figure 1.

- **Event semantic decoupling** refers to the requirement that event producers and consumers can exchange events without a-priori full understanding of, or agreement on the vocabulary used to describe event types, properties or values.

3.1 Current Approaches

Most work in event-based systems assumes well-defined and agreed upon semantics of events. Work focusing on event interoperability such as S-TOPSS [24], FOMatch [29] or the Open Geospatial Consortium's Sensor Web Enablement (SWE) [6] target semantic heterogeneity using common vocabularies/ontologies or other kinds of representation of domains in order to address large scale event heterogeneity. However, tying participants to definitions of domains adds a higher-level explicit dependency between them. It also limits new participants with new event semantics that are not covered by the common vocabularies from joining the system. In other words, this approach addresses semantic heterogeneity but not semantic decoupling and as a result, it does not scale out to large scales of heterogeneity. An analogous experience is found in the semantic web community where generating consensus on common ontologies is known to be very challenging [11]. Another work, A-TOPSS [17], targets coupling on value level by enabling imprecise subscription with fuzzy membership functions of numeric event values. However, type, property and non-numeric event values are not considered.

3.2 Contribution

The contribution of this paper is threefold:

- An approach to address semantic decoupling via approximate semantic matching of events is proposed.
- A hybrid instantiation of an approximate semantic event matcher based on both thesauri and distributional semantics-based semantic similarity and relatedness measures is presented.
- An evaluation of the approach based on real world events from Wikipedia and Freebase with a set of gold standard subscriptions is discussed.

4. PROPOSED APPROACH

In this paper we propose an approximate semantic matching approach to semantic decoupling. The approach builds on the semantic decoupling principle and proposes an event model, a subscription model and a matching model that leverage semantics of events and subscriptions to establish approximate matching as illustrated by Figure 2. The following subsections discuss these various models and the challenges associated with the matching steps.

4.1 Event Model

The proposed event model frees event producers from using agreed-upon vocabulary to describe events. Each event is an instance of one or more event types $\{T_1, T_2,\dots, T_l\}$. It has one or more event properties $\{P_1, P_2,\dots, P_m\}$. It may have one or more values for each event property $\{(P_1, v_1), (P_2, v_2),\dots, (P_{m},, v_n)\}$. Terms used for types, properties or values do not need to conform to an agreed-upon vocabulary with other parties in the system. Example 2 presents an event example

Example 2

The *2010 FIFA World Cup Final* is an event which has the type *"Football Match"*, a set of properties {*"name", "referee", "location", "team"*}, and a set of values associated with the properties {*("name", "2010 FIFA World Cup Final"), ("referee", "Howard Webb"), ("location", "FNB Stadium"), ("location", "Johannesburg"), ("team", "Spain national football team"), ("team", "Netherlands national football team")*}.

4.2 Subscription Language Model

The proposed subscription model frees event consumers from using predefined vocabulary to define subscriptions. A subscription describe a set of one or more types of interest $\{T_{S1'}, T_{S2'},\dots, T_{Sl'}\}$. It specifies one or more properties to filter on $\{P_{S1'}, P_{S2'},\dots, P_{Sm'}\}$. It specifies one or more values for equality check for each property $\{(P_{S1'}, v_{S1'}), (P_{S2'}, v_{S2'}),\dots, (P_{Sm'}, v_{Sn'})\}$. Example 3 shows a subscription meant to match the event in Example 2.

Example 3

An example subscription is a subscription interested in events with the type *"Soccer Match"*, a set of properties {*"place", "team"*}, and a set of values associated with these properties {*("place", "South Africa"), ("team", "Spain")*}.

4.3 Matching Model

The core principle of our proposed approach is based on the removal of explicit semantic dependencies between interacting participants in an event-based system. The main steps as illustrated in Figure 2 are:

1. **Establish approximate mappings between the type(s) & properties of an event, to the type(s) & properties of the subscription.** This may be established based on semantic interpretation with a quantification that reflects the degree of approximation for types and properties in the mapping. E.g. given the event in Example 2 and the subscription in Example 3, possible mappings may be: $Mapping_1$ = {(*"Football Match"* ↔ *"Soccer Match"*), (*"location"* ↔ *"place"*), (*"team"* ↔ *"team"*)} with 90% quantification, $Mapping_2$ = {(*"Football Match"* ↔ *"Soccer Match"*), (*"referee"* ↔ *"team"*), (*"name"* ↔ *"team"*)} with 40% quantification, etc.
2. **Establish approximate semantic mapping between values in the event and corresponding values in the subscription.** The matching results in a quantification that reflects the degree of semantic approximation between values. E.g. in $Mapping_1$ given above and for the property *"team"*, an approximate semantic matching (*"Spain national football team"* ↔ *"Spain"*) would be established with 95%

quantification rather than the matching (*"Netherlands national football team"* ↔ *"Spain"*) of 45% quantification.

3. **Use the best overall mapping** of types, properties and values as the correct event-to-subscription mapping. The result of matching is the overall quantification score associated with that mapping. E.g. *Mapping$_1$* mentioned earlier with value matching (*"Johannesburg"* ↔ *"South Africa"*) and (*"Spain national football team"* ↔ *"Spain"*) would be considered the best with an overall quantification of 80%.
4. **Post-matching event processing**. This can include cutting off events with a matching score less than a threshold, etc. E.g. the previous matching of the event in Example 2 and the subscription in Example 3 with a matching score of 80% would be passed through a filter of a 75% threshold.

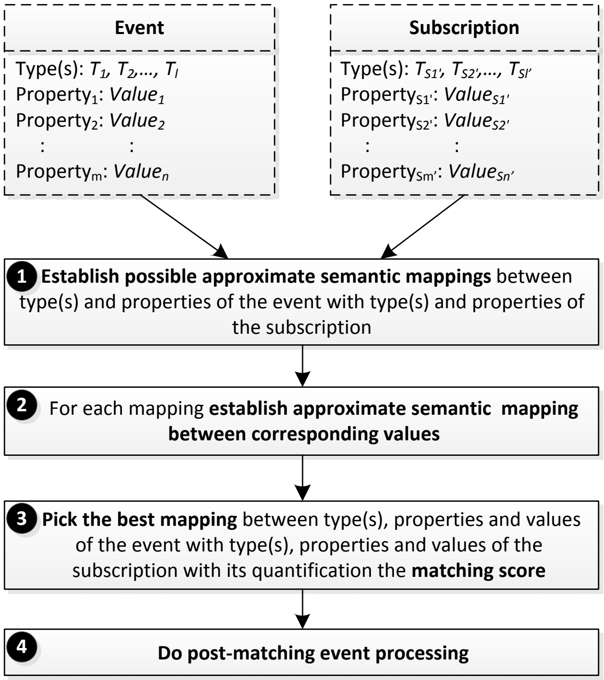
Event
Type(s): $T_1, T_2, \ldots, T_l$
Property$_1$: $Value_1$
Property$_2$: $Value_2$
: :
Property$_m$: $Value_n$
Subscription
Type(s): $T_{S1'}, T_{S2'}, \ldots, T_{Sl'}$
Property$_{S1'}$: $Value_{S1'}$
Property$_{S2'}$: $Value_{S2'}$
: :
Property$_{Sm'}$: $Value_{Sn'}$
1 Establish possible approximate semantic mappings between type(s) and properties of the event with type(s) and properties of the subscription
2 For each mapping establish approximate semantic mapping between corresponding values
3 Pick the best mapping between type(s), properties and values of the event with type(s), properties and values of the subscription with its quantification the matching score
4 Do post-matching event processing

Figure 2. Proposed approach

4.4 Challenges

In order to deliver an approximate event matcher there are a number of research questions which need to be answered in the proposed matching model:

Approximate Semantic Matching of Types, Properties and Values

What is the most suitable mechanism to quantify semantic matching of terms used to describe types, properties and values of events? The mechanism must not assume existing semantic coupling between event producers and consumers.

Ranking Possible Mappings

What is the most suitable way to combine quantifications of types, properties and values mappings so the overall mappings between an event and a subscription can be ranked?

Handling of Background Knowledge

Background knowledge may be needed in order to quantify approximate semantic matching of terms. For example, a football match event in *"Johannesburg"* would match a subscription for football matches in *"South Africa"* only if the background knowledge that *"Johannesburg is a city in South Africa"* is available.

Handling of Uncertain Matching

The result of approximate matching is not Boolean but rather a score of approximation required to the matching. This score reflects the uncertainty embedded in the event-to-subscription matching decision. Uncertainty must be interpreted in a way that allows it to be propagated for further post-matching phases such as combining scores from several matching events to evaluate a pattern of events.

5. INSTANTIATION

The proposed approach for event and subscription models is instantiated using a triple-based framework. The matching model is realized by leveraging semantic similarity and relatedness functions to address challenges discussed in the previous section.

5.1 Event Model

An event is a labelled directed graph as illustrated in Figure 3. The Resource Description Framework (RDF) [16] is used to represent information about events using statements. A statement consist of a (subject, property, object) triple. Properties may come from various vocabularies (the RDF name of ontologies) and one subject may have multiple statements with the same property and different objects. In this paper, the following restrictions apply to an RDF event graph:

- All the statements of the event graph share the same reference subject. The subject resource represents the event.
- An event type is represented using a statement with an appropriate property which is semantically similar to "*type*" and a suitable reference to a vocabulary class for the object. An event may have more than one type statement.

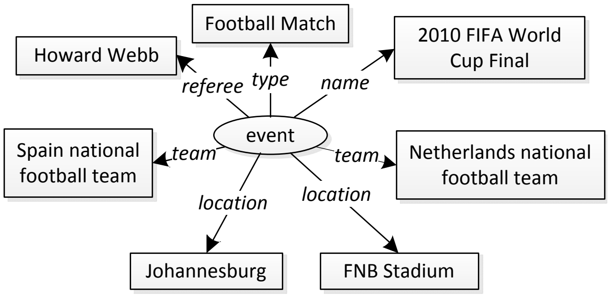

Figure 3. An example event.

The resulting event can be represented as follows: Let *E* be the set of events conforming to the event model, *S* the set of subjects, *P* the set of properties and *V* the set of values, then an event can be seen as a finite set of triples with the same subject and with the attention to the order of the parts of a triple, as follows:

- $e \in E \Leftrightarrow e = \{(s,p,v) : (s,p,v) \in S \times P \times V$ $\wedge \forall (s_1,p_1,v_1),(s_2,p_2,v_2) \in e \Rightarrow s_1 = s_2\}$ (1)
- The set of types of *e* is *T*= $\{t : (s,p,t) \in S \times P \times V$ $\wedge\ p$ is semantically similar to "*type*"} (2)

5.2 Subscription Language Model

As the event model is a set of triples, the subscription model is a finite set of conjunctive statements. The property and object parts of a subscription triple are free natural English language. The filtering operator is the equality operator with the relaxation of equality of terms to approximate semantic matching of terms. The operator is not explicitly expressed in the language but the existence of a particular object value in the triple means interest in an event with an object value that semantically matches the subscription one. Types can be expressed in the same way other properties are expressed using a property that semantically matches "*type*". The resulting language model is as follows: Let *Sub* be the set of subscriptions conforming to the subscription model, *S* the set of subjects, *P* the set of properties and *V* the set of values, then:

- $s \in Sub \Leftrightarrow s = \{ (s, p, v) : (s, p, v) \in S \times P \times V \}$ with the semantic of a conjunction between statements. (3)

Example 4 illustrates a subscription intended to match the event represented in Figure 3.

Example 4: An example subscription

event	*type*	*"Football Match"*
event	*team*	*"Spain national football team"*
event	*stadium*	*"FNB Stadium"*

5.3 Matching Model

5.3.1 Concepts and Definitions

The proposed matching model instantiation is based on the notion of relatedness between an event *e* and a subscription *s*. The result of a match is not a crisp Boolean value but rather a real number that reflects the degree of relatedness of *e* to *s*. Semantic matching builds on the assumption that an event *e* and a subscription *s* are more related if they share more semantically similar or related correspondent types, properties and values. In the following, relatedness is considered as a generic case of similarity. Before we instantiate the steps of the proposed matching model presented in Figure 2, the following concepts are defined:

Relatedness Functions

Given two terms t_1 and t_2 of the English language term set *TERMS*, the relatedness between the two terms values is quantified by means of a relatedness function *f* which is defined as follows:

$$f : TERMS \times TERMS \rightarrow [0,1] \qquad (4)$$

f is a symmetric function and returns the value 1 for an exact match, 0 for an absolute mismatch and another value from the range (0, 1) otherwise. f_V denotes a relatedness function defined over values. f_P denotes a relatedness function defined over properties. Types are handled similarly to values by f_V as the proposed event model instantiation suggests the use of triples for both values and types.

Statement Relatedness

Given a subscription statement $stmt_s$ = (*event*, p_s, v_s) and an event statement $stmt_e$ = (*event*, p_e, v_e), the relatedness between the two statements is quantified by means of a relatedness function f_{STMT} which is defined as follows:

$$f_{STMT} : (S \times P \times V) \times (S \times P \times V) \rightarrow [0,1] \qquad (5)$$

$$f_{STMT}(stmt_s, stmt_e) = f_P(p_s, p_e) * f_V(v_s, v_e) \qquad (6)$$

Statement Correspondence

Given an event *e* and a subscription *s*, an event statement $stmt_e$ is called correspondent to a subscription statement $stmt_s$, denoted as $stmt_e = Corr_e(stmt_s)$, if the statement relatedness between $stmt_s$ and $stmt_e$ is higher than or equal to the statement relatedness between $stmt_s$ and any other statement of *e* other than $stmt_e$. $Corr_e$ is thus defined as follows:

$$Corr_e : s \rightarrow e \qquad (7)$$

$$stmt_e = Corr_e(stmt_s) \Leftrightarrow \forall stmt_{ei} \in e \wedge stmt_{ei} \neq stmt_e \qquad (8)$$
$$\Rightarrow f_{STMT}(stmt_s, stmt_e) \geq f_{STMT}(stmt_s, stmt_{ei})$$

T*op*$_m$ Mapping Candidates

Given an event *e*, a subscription *s* and a subscription statement $stmt_s$= (*event*, p_s, v_s) $\in s$, a list of *m* event properties is called the Top_m correspondence candidates list for p_s if the property of $Corr_e(stmt_s) \in Top_m$.

$$Top_m : P \rightarrow 2^P \qquad (9)$$

Ranking Function

This function ranks the possible overall mappings of types, properties and values between an event *e* and a subscription *s*. It also ranks different events according to their semantic relatedness to the subscription. The model instantiation defines a relatedness function f_R as the mean of relatedness values of each subscription statement and its correspondent as follows:

$$f_R : E \times S \rightarrow [0,1] \qquad (10)$$

$$f_R(e,s) = \frac{\sum_{i=1}^{|s|} f_{STMT}(stmt_{si}, Corr_e(stmt_{si}))}{|s|} \qquad (11)$$

Where $stmt_{si}$ is a statement in the subscription *s* and |*s*| is the number of statements in the subscription *s*. The ranking function orders the events according to their descending relatedness values to a subscription *s*.

Event Match

Given an event *e*, a subscription *s* and a *threshold* function defined over the subscription set as follows:

$$threshold : S \rightarrow [0,1] \qquad (12)$$

Event *e* matches subscription *s* iff

$$f(e,s) \geq threshold(s) \qquad (13)$$

The *threshold* function can be a constant value for all the subscriptions or provided by the user. The Boolean matching model is given as an example of post-processing of approximate semantic matching. For the rest of this paper the *threshold* function is discarded and more emphasis is given to ranking during the evaluation.

5.3.2 Matching Model Instantiation

Instantiation of the matching model requires a proper combination of the functions defined in the previous section to realize the steps of the matching model illustrated in Figure 2. It also requires a suitable instantiation of the value relatedness function f_V, the property relatedness function f_P and the Top_m mapping candidates function. This section firstly discusses the steps and then moves to the realization of functions.

Matching Steps

Figure 4 illustrates actual instantiation of matching steps as follows:

1. To establish possible approximate semantic mappings, the Top_m mapping candidates function is used to suggest m possible mapping from the event properties for each subscription property (a). The property relatedness function f_P quantifies each possible mapping by quantifying the different pairs of property mappings (b).
2. Approximate semantic matching between values is achieved using the value relatedness function f_V (c).
3. The best overall mapping is achieved by means of the relatedness function f_{STMT} (d), the statement correspondence function $Corr_e$ (e), and the ranking function f_R (f).
4. Post-matching can be realized using the ranking function f_R itself to rank different events according to the same subscription. Boolean event matching using a *threshold* function is another type of post-matching using approximate model (g).

Value Relatedness

In order to instantiate the function f_V for a domain-agnostic purpose, an appropriate semantic similarity or relatedness function should be used. Existing research [12] has proved superiority of semantic relatedness measures based on distributional semantics of words in large open domain corpora such as Wikipedia. Distributional semantics measures typically handle single or multi-word terms in different morphological forms or word classes and support semantic relatedness which is a generic case of semantic similarity. E.g. "football match" and "soccer match" are similar in terms of the synonymy relationship while *"football match"* and *"referee"* are not similar but still related terms.

The fact that distributional semantics measures are typically built over large common knowledge corpora implies that implicit background knowledge is encoded in these measures. E.g. the pair (*"FNB Stadium"*, *"Johannesburg"*) scores relatively high in a Wikipedia-based measure due to existence of background knowledge that *FNB Stadium* exists in an area of *Johannesburg*. Examples of such background knowledge are the Wikipedia pages about *FNB Stadium (Soccer City)* and *Johannesburg*.

The proposed model instantiation uses the Wikipedia-based Explicit Semantic Analysis (ESA) relatedness measure, denoted as *WikipediaESA* to realize the function f_V. The main strong points of Wikipedia-based ESA come from the fact that Wikipedia is one of the most comprehensive and high quality world knowledge base created by human. Wikipedia-based ESA represents each word by a vector of Wikipedia articles' titles that are based on a TF/IDF weighing scheme. Semantic relatedness between two words is computed using the cosine metric of the two words' vectors [12]. Wikipedia-based ESA assigns scores normalized between 0 and 1.

Property Relatedness

One possible instantiation of the f_P function is to use Wikipedia-based ESA as it is used for value relatedness. However, properties in events are more precise and controlled; hence they may come from well defined ontologies. Properties as used in subscriptions are also more specific than values. Thus, measuring relatedness between properties does not significantly benefit from the strengths of Wikipedia-based ESA, namely multi-word terms, handling different morphological forms and word classes, and implicit background knowledge. Empirical results of the first experiment presented in section 6.1 of this paper confirmed the aforementioned observation when using ESA to realize f_P for single-event matching purpose as shown in Figure 5.

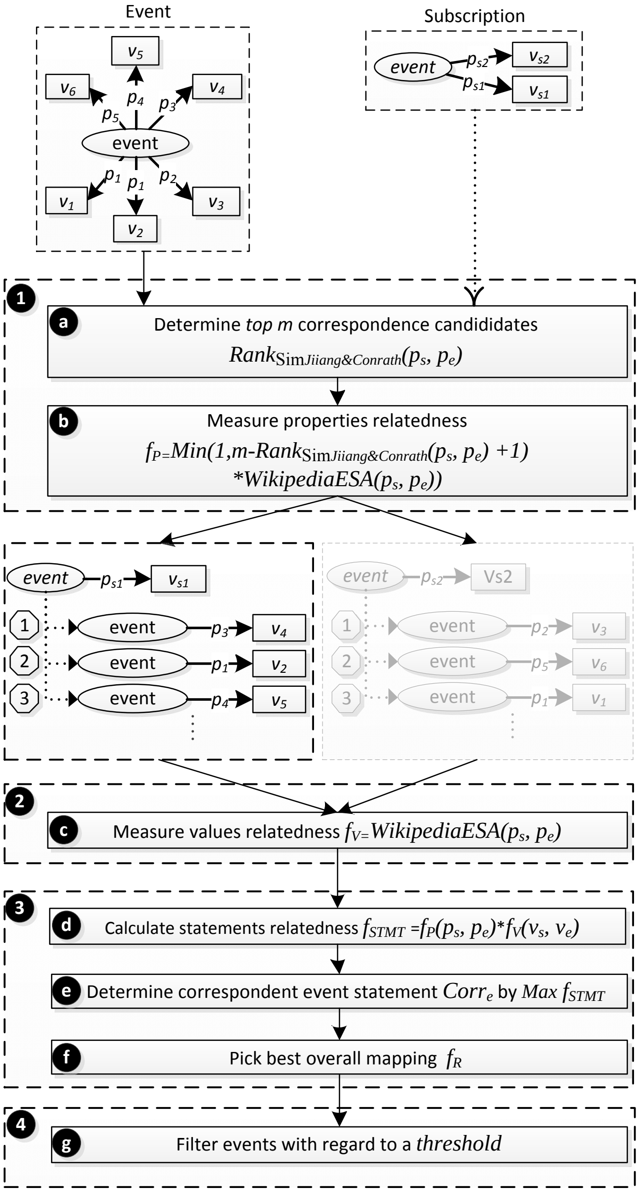
Event
Subscription
v_5
v_6
v_4
p_4
p_3
p_5
event
p_1
p_1
p_2
v_1
v_2
v_3
event
p_{s2}
v_{s2}
p_{s1}
v_{s1}
1
a
Determine *top m* correspondence candididates
$Rank_{\text{Sim}_{Jiiang\&Conrath}}(p_s, p_e)$
b
Measure properties relatedness
$f_{P=}Min(1,m\text{-}Rank_{\text{Sim}_{Jiiang\&Conrath}}(p_s, p_e) +1)$
$*WikipediaESA(p_s, p_e))$
event
p_{s1}
v_{s1}
1
event
p_3
v_4
2
event
p_1
v_2
3
event
p_4
v_5
event
p_{s2}
Vs2
1
event
p_2
v_3
2
event
p_5
v_6
3
event
p_1
v_1
2
c
Measure values relatedness $f_{V=}WikipediaESA(p_s, p_e)$
3
d
Calculate statements relatedness $f_{STMT}=f_P(p_s, p_e)*f_V(v_s, v_e)$
e
Determine correspondent event statement $Corr_e$ by *Max* f_{STMT}
f
Pick best overall mapping f_R
4
g
Filter events with regard to a *threshold*

Figure 4. Matching model instantiation

A more appropriate choice to instantiate f_P is to use a semantic similarity measure that is based on human created and domain-agnostic thesaurus. WordNet-based similarity measures are good candidates for this as WordNet is one of the largest structured English language thesauri maintained by humans [21]. WordNet-based measures can be classified into:

- **Edge counting-based measures** such as the shortest path [27] where the shorter the path between two concepts in WordNet the more similar the concepts.

- **Information content-based measures** such as Jiang&Conrath [15] where information content of a concept increases as it becomes more specific, i.e. closer to leaves in WordNet. The measure exploits th**e information content of the share**d parents of two concepts as well as the concepts themselves.
- **Feature-based measures** such as gloss vectors [23] where word definitions of the WordNet terms are compared using a cosine measure.

Previous research by Budanitsky and Hirst [7] as well as empirical results based on the first experiment presented later in this paper, as illustrated by Figure 5, show superiority of the Jiang&Conrath [15] information content-based similarity measure. The Jiang&Conrath semantic similarity function denoted as $Sim_{Jiang\&Conrath}$ measures semantic similarity between WordNet synonyms sets (synsets) and may give arbitrarily large values. To compute the measure for single-word terms, the most common synset is used. For multi-word terms, the mean of the maximum words pair-wise scores is used. Scores are then normalized to the range [0, 1] with the 1 representing an exact match.

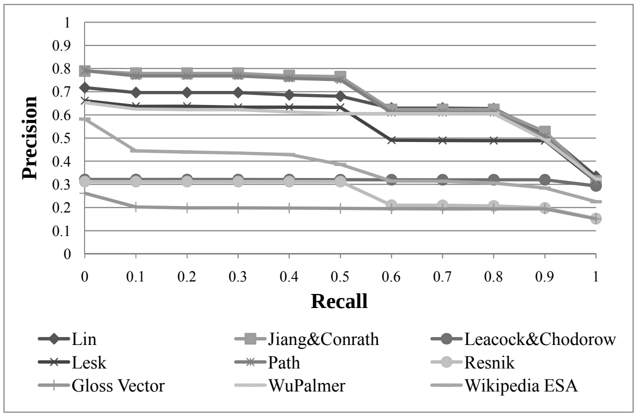

Figure 5. Comparing different similarity and relatedness measures when the same measure is used to define f_P and f_V in the Wikipedia experiment. Refer to [7] **and** [12] **for information on various measures.**

Nevertheless, using $Sim_{Jiang\&Conrath}$ to realize f_P causes a problem when combining it with *WikipediaESA* to compute the statement relatedness as in Equation (6). This is because the two measures values have quite different distributions as shown by Figure 6. To explain this, assume that terms t_1, t_2, t_3. $t_4 \in V$ can be used as properties or values with $t_1 = t_2$ and t_3, t_4 are similar but not identical. The pair of statements (*event*, t_1, t_3) and (*event*, t_2, t_4) would result in a much higher statement relatedness score than the pair (*event*, t_3, t_1) and (*event*, t_4, t_2) according to Equation (6) which is counter-intuitive.

In order to address these issue, we chose to use *WikipediaESA* for property relatedness which guarantees the homogeneity of f_P and f_V distributions. We propose the use of a corrective factor for *WikipediaESA* when used for f_P to benefit from the strengths of the Jiang&Conrath similarity measure and overcome the challenges associated with *WikipediaESA* for property relatedness. To define the corrective factor, we let $Rank_{\mathrm{Sim}Jiiang\&Conrath}$ be the ranking function of event properties to a subscription property according to $Sim_{Jiang\&Conrath}$, the final instantiation of the property relatedness function f_P employs the inverse rank of the event property to weigh the *WikipediaESA* measure as follows:

$$f_P(p_s, p_e) = Min(1, (m - Rank_{Sim\,Jiang\&Conrath}(p_s, p_e) + 1) \\ *WikipediaESA(p_s, p_e)) \qquad (14)$$

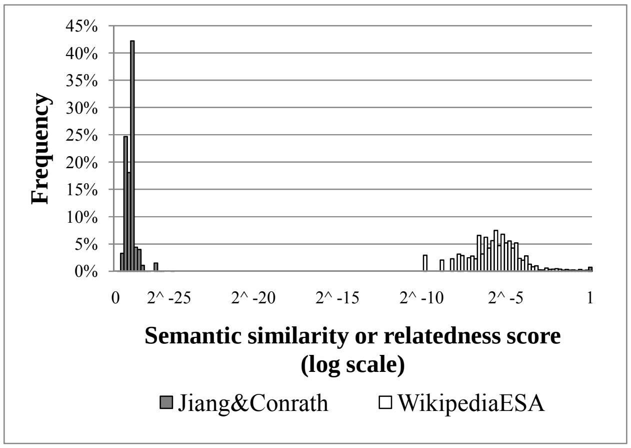

Figure 6. Histogram of f_P = $Sim_{Jiang\&Conrath}$ and f_V=*WikipediaESA* in a sample based on matching DBPedia events in experiment 1. Hence both functions give value 1 for some pairs.

The strategy of composing various measures and matching methods has been successfully applied in the context of schema and ontology matching [3].

Top$_m$ Correspondence Candidates

The proposed Top_m function picks the event statements with the highest property relatedness to the property of subscription statement based on the $Sim_{Jiang\&Conrath}$ score of similarity. Given an event *e*, a subscription *s*, a subscription statement $stmt_s$ = (*event*, p_s, v_s) $\in s$, $Rank_{\mathrm{Sim}Jiiang\&Conrath}$ the ranking function of event properties according to $Sim_{Jiang\&Conrath}$, Top_m is defined as follows:

$$Top_m(stmt_s) = \{(event, p_e, v_e) : (event, p_e, v_e) \in e \\ \wedge Rank_{Sim\,Jiang\&Conrath}(p_s, p_e) \le m\} \qquad (15)$$

Results presented in the paper are for a value $m = 5$ unless mentioned otherwise. Figure 4 illustrates the proposed model instantiation.

6. EXPERIMENTS

To evaluate the approximate semantic matching approach of events, we have chosen to leverage online datasets, namely Wikipedia and Freebase, as an experimental basis. That is due to the fact that they reflect the high heterogeneity and the need for semantic decoupling that this paper tackles. A set of gold standard subscriptions was prepared

6.1 Wikipedia Experiment

Event Set

Wikipedia[1] is one of the most comprehensive common knowledge base created by a distributed and decoupled community of human contributors. This makes Wikipedia a very

[1] http://www.wikipedia.org/

good resource to emphasize the semantic heterogeneity for event systems. Wikipedia provides a natural language document based encyclopaedia that describes different types of entities: people, places, events, etc. The event set used is a structured representation of events in Wikipedia. DBpedia [2] is a community project to extract structured information from Wikipedia. DBpedia is one of the efforts under the Linked Open Data initiative which targets the publication of structured data on the web according to the Linked Data principles [5]. The data model used to represent DBPedia data is RDF. The event set used for this experiment is a subset of the current version the English DBPedia[1]. It contains all the resources of type `dbpedia-owl:Event` and their directly associated triples. Table 1 summarizes the characteristics of the event set.

Table 1. Characteristics of the Wikipedia event set

Source and Collection	
English DBpedia current version, last modified on the 31st of August 2011. All triples directly associated with instances of class `dbpedia-owl:Event`	
Statistics	
Data model	RDF
Total # of events	~ 20,000
Total # of distinct event types	~ 4,950
Total # of distinct event properties	~ 1,500
Total # of distinct event values	~ 500,000
Total # of triples	~ 1,500,000
Average # of distinct type per event	~ 7
Average # of distinct property per event	~ 31
Average # of distinct value per event	~ 54
Average # of triples per event	~ 65

Events are pre-processed before matching to conform to the event model detailed in Section 5.1. URIs for example are manipulated to extract the last part after a "/" or a "#". Underscores are removed from strings and camel case strings are separated into words. Nouns are singularized and values with RDF language tags other than English are discarded. Example event types that can be found in the event set are: *"Football Match"*, *"Race"*, *"Music Festival"*, *"Space Mission"*, *"Election"*, *"10th-century BC Conflicts"*, *"Academic Conferences"*, *"Aviation Accidents And Incidents In 2001"*, etc.

Subscription Set

A set of 7 subscriptions has been prepared manually. Each subscription uses free English terms for properties and values. Subscriptions stress the type, property, and value dimensions of semantic decoupling in order to be a representative set of possible subscriptions. In order to specify the set of events that is relevant to each subscription, each subscription was translated manually to one or more exact SPARQL [26] patterns[2]. SPARQL is an exact query language for RDF. The union of resulting events that exactly match the SPARQL queries is considered the set of relevant events for the subscription in question. For example, the subscription that represents the information need *"Events taking place in Wembley stadium"* is represented according to the subscription model as in Table 2. The exact matching SPARQL patterns for this subscription are also shown in Table 2.

Table 2. Subscription and translation for "Events taking place in Wembley stadium"

Subscription	event type "Event" event place "Wembley Stadium"

SPARQL pattern 1	`?event rdf:type dbpedia-owl:Event.`
	`?event dbpprop:stadium`
	`dbpedia:Wembley_Stadium.`

SPARQL pattern 2	`?event rdf:type dbpedia-owl:Event.`
	`?event dbpedia-owl:location`
	`dbpedia:Wembley_Stadium.`

SPARQL pattern 3	`?event rdf:type dbpedia-owl:Event.`
	`?event dbpedia-owl:location`
	`dbpedia:Wembley_Stadium_(1923).`

SPARQL pattern 4	`?event rdf:type dbpedia-owl:Event.`
	`?event dbpprop:stadium`
	`dbpedia:Wembley_Stadium_(1923).`

SPARQL pattern 5	`?event rdf:type dbpedia-owl:Event.`
	`?event dbpprop:venue`
	`dbpedia:Wembley_Stadium_(1923).`

Table 3 illustrates the 7 subscriptions and their corresponding information needs.

Table 3. The 7 subscriptions used for evaluation

ID	**Information Need**	**Subscription**
1	Football matches played by Spain in the FNB stadium	event type "Football Match" event team "Spain national football team" event stadium "FNB Stadium"
2	Football matches played in the FNB stadium	event type "Football Match" event place "FNB Stadium"
3	Events taking place in Wembley stadium	event type "Event" event place "Wembley Stadium"
4	Charity events taking place in Wembley stadium	event type "Charity" event place "Wembley Stadium"
5	Charity Rock events taking place in Wembley stadium	event type "Charity" event type "Rock" event place "Wembley Stadium"
6	Football matches played in the UK	event type "Football Match" event stadium "United Kingdom"
7	Football matches played by a South American team in Europe	event type "Football Match" event team "South America" event stadium "Europe"

Table 4 shows the number of relevant events for each subscription and the number of exact matching patterns that would be needed in order to match the whole set of relevant events. A classification of the subscriptions according to their coverage of the semantic decoupling dimensions is also presented. "BK" refers to the need for background knowledge to accomplish the matching.

Table 4. Characteristics of the subscription set with regard to Wikipedia event set

ID	**# of Relevant Events**	**# of Needed Exact Patterns**	**Approximation**		
			Type	**Property**	**Value**
1	1	1	No	No	No
2	2	2	No	Yes	No
3	219	5	No	Yes	No
4	29	6	Yes	Yes	Yes
5	2	2	Yes	Yes	Yes
6	505	603	No	Yes	BK
7	20	123,774	No	Yes	BK

[1] http://downloads.dbpedia.org/current/en/. Last modified on the 31st of August 2011. Accessed on 30th of May 2012.

[2] A SPARQL pattern is a pattern of statements meant to match RDF triples for one event. It is not to be confused with an event pattern which is meant to match a set of events.

Evaluation

The adopted evaluation is based on the framework used to evaluate approaches in content-based information filtering, a field concerned mainly with approximate matching of documents to user interests [4]. The framework in turn is based on the one traditionally used in information retrieval [20].

The main evaluation criteria are precision, recall and the combined metric F_1 Score. These are defined as follows:

$$Precision = \frac{\#(relevant\ events\ matched)}{\#(matched\ events)} \quad (16)$$

$$Recall = \frac{\#(relevant\ events\ matched)}{\#(relevant\ events)} \quad (17)$$

$$F_1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (18)$$

To evaluate the ranking behaviour of the hybrid matcher, the 11-point interpolated average precision-recall curve is used [20]. For each subscription, the interpolated precision is measured at 11 recall levels of 0.0, 0.1, 0.2, …, 1.0 where interpolated precision is the maximum precision value that could be got for all the recall levels greater than or equal to the recall level in question. The interpolated precision measures of all the subscriptions are averaged then on all the recall levels to give the overall matcher curve.

The proposed hybrid approximate semantic matcher, which combines $Sim_{Jiang\&Conrath}$ and *WikipediaESA* for instantiating f_P, f_V and Top_m functions, is compared with two matchers representing two classes of possible instantiations of the proposed matching model in Section 5.2:

Jiang&Conrath matcher which represents WordNet-based matchers and instantiates the matching model as follows:

$$f_P(p_s, p_e) = Sim_{Jiang\&Conrath}(p_s, p_e) \quad (19)$$

$$f_V(v_s, v_e) = Sim_{Jiang\&Conrath}(v_s, v_e) \quad (20)$$

$$Top_m(stmt_s) = \{(event, p_e, v_e) : (event, p_e, v_e) \in e \wedge Rank_{Sim_{Jiang\&Conrath}}(p_s, p_e) \leq m\} \quad (21)$$

Wikipedia ESA matcher which represents Wikipedia-based distributional semantic matchers and instantiates the matching model as follows:

$$f_V(v_s, v_e) = WikipediaESA(v_s, v_e) \quad (22)$$

$$f_P(p_s, p_e) = WikipediaESA(p_s, p_e) \quad (23)$$

$$Top_m(stmt_s) = \{(event, p_e, v_e) : (event, p_e, v_e) \in e \wedge Rank_{WikipediaESA}(p_s, p_e) \leq m\} \quad (24)$$

Figure 7 illustrates the resulting average interpolated precision-recall curves for the three matchers. It shows that the proposed hybrid matcher performs better than the other two along all the recall levels. Wikipedia ESA matcher performs badly with subscriptions that stress event type or properties approximation; Figure 8 compares the matchers' effectiveness with respect to Subscription 3. Conversely, Jiang&Conrath fails to respond to subscriptions that stress value approximation especially with need for background knowledge; Figure 9 compares the matchers' effectiveness with respect to Subscription 6. This observation validates the approach of combining the two measures in a way that the resulting matcher can benefit from the strengths of both of them.

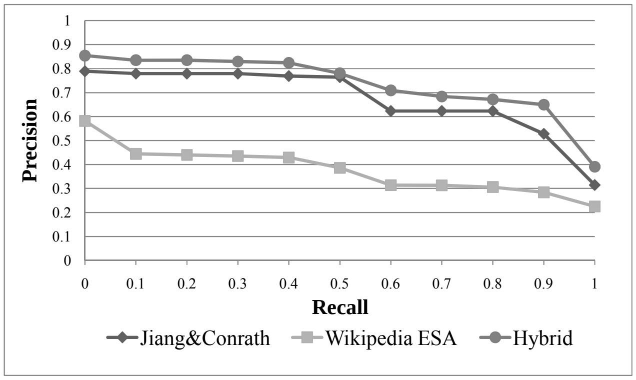

Figure 7. Interpolated average precision-recall curve for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher.

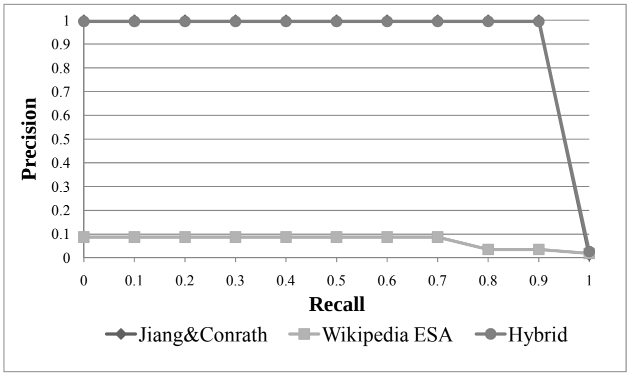

Figure 8. Interpolated precision-recall curve for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher for Subscription 3 for Wikipedia event set. Curves of Jiang&Conrath and Hybrid matchers are overlapped.

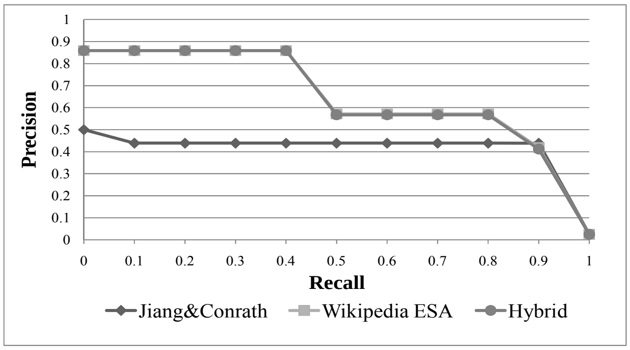

Figure 9. Interpolated precision-recall curve for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher for Subscription 6 for Wikipedia event set. Curves of Wikipedia ESA and Hybrid matchers are overlapped.

The F_1 Score combines the precision and recall measures in one metric. Thus, finding the Maximal F_1 Score along a matcher's precision-recall curve helps summarizing the overall matcher's effectiveness in one number. Table 5 shows the Maximal F_1 Score for the three matchers. The proposed hybrid matcher outperforms the other two as suggested previously by the interpolated average precision-recall curves in Figure 7.

Table 5. Maximal F_1 Score for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher for the Wikipedia event set

Matcher	Jiang&Conrath	Wikipedia ESA	Hybrid
Maximal F_1 Score	70.06%	44.26%	75.45%
Recall	80%	80%	90%
Precision	62.31%	30.59%	64.94%

Another dimension to study the proposed approximate semantic matching model and the hybrid instantiation is by examining the effect the parameter *m* in the Top_m function has on the interpolated precision-recall curve. Figure 10 shows this effect and suggests that the overall effectiveness increases as *m* increases. However, the improvement that increasing *m* causes quickly becomes small and a low number of *m* can provide effectiveness very close to the maximum.

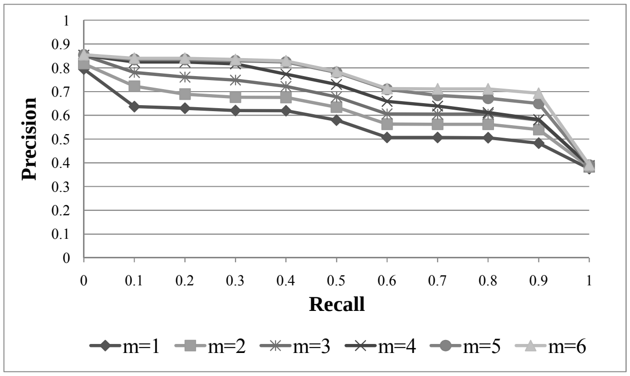

Figure 10. Effect of *m* in the Top_m function on the overall effectiveness of the proposed hybrid matcher for Wikipedia event set.

6.2 Freebase Experiment

The second experiment aims at evaluating the effectiveness of the proposed approach when a new event source with new event types, properties and values is added while the set of subscriptions is kept un-touched.

Event Set

Freebase[1] is a community driven repository of structured data on the web. Entities in Freebase contain people, places, events, etc. The event set used for this experiment is a subset of Freebase containing all instances of type `fbase:time.event`. Instances were retrieved from a Freebase dump[2]. Triples that are directly associated with each event instance are retrieved by dereferencing its URI via the Freebase RDF service[3]. Events are pre-processed before matching to conform to the event model in the same way as in the first experiment. Table 6 summarizes the characteristics of the event set.

Table 6. Characteristics of the Freebase event set

Source and Collection	
Freebase dump, last modified on the 1st of December 2011. All triples directly associated with instances of class `fbase:time.event`	
Statistics	
Data model	RDF
Total # of events	~ 85,000
Total # of distinct event types	~ 850
Total # of distinct event properties	~ 1,250
Total # of distinct event values	~ 1,200,000
Total # of triples	~ 1,850,000
Average # of distinct type per event	~ 3
Average # of distinct property per event	~ 11
Average # of distinct value per event	~ 22
Average # of triples per event	~ 22

Subscription Set

The set of subscriptions is the same used in the previous experiment and presented in Table 3. Determining the relevant events to each subscription is done in the same manner. However, no relevant events were found for subscriptions 4 and 5 so they are excluded. Table 7 characterizes the set of subscriptions with regard to the Freebase event set.

Table 7. Characteristics of the subscription set with regard to Freebase event set

ID	# of Relevant Events	# of Needed Exact Patterns	Approximation		
			Type	Property	Value
1	1	1	Yes	Yes	No
2	8	2	Yes	Yes	No
3	29	5	No	Yes	No
4	0	-	-	-	-
5	0	-	-	-	-
6	34	1,398	Yes	Yes	BK
7	2	219,600	Yes	Yes	BK

Evaluation

The same evaluation criteria from the previous experiment were used, i.e. Precision, Recall and the Maximal F_1 Score. Figure 11 illustrates the resulting average interpolated precision-recall curves for the three matchers: Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher. It shows that the proposed hybrid matcher performs better, confirming the results of the first experiment.

Wikipedia and Freebase events are open domain event sets where terms are not biased toward a domain specific distribution. *WikipediaESA* relatedness measure in turn is based on an open domain corpus which is Wikipedia. Experiments thus corroborated that *WikipediaESA* does not bias toward the first experiment where events themselves come from Wikipedia.

[1] http://www.freebase.com/

[2] http://download.freebase.com/datadumps/. Dump retrieved on the 1st of December 2011.

[3] http://rdf.freebase.com/

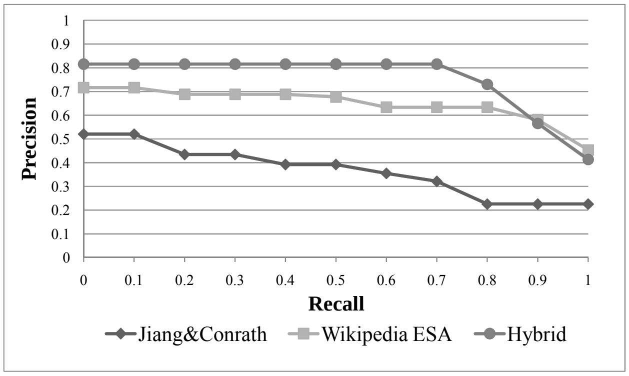

Figure 11. Interpolated average precision-recall curve for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher for Freebase event set.

Table 8 summarizes the matchers' effectiveness using the Maximal F_1 Score. Results show the effectiveness of the hybrid matcher for both Wikipedia and Freebase event sets. However, no cost is associated with defining new subscriptions to match the Freebase events. The semantic matcher is able to process these new events using the same subscriptions. That indicates the suitability of the proposed approach within heterogeneous environments. Across both event sets the matcher has effectively replaced almost 345,000 needed exact matching patterns: 125,000 of which for Wikipedia events and 220,000 for Freebase events.

Table 8. Maximal F_1 Score for Jiang&Conrath, Wikipedia ESA and the proposed Hybrid matcher for the Freebase event set

Matcher	Jiang&Conrath	Wikipedia ESA	Hybrid
Maximal F_1 Score	44.60%	70.73%	76.33%
Recall	60%	80%	80%
Precision	35.49%	63.39%	72.98%

7. RELATED WORK

Approximate Event Matching

One of the early work on approximate event matching is A-TOPSS [17] which defines an approximate matching model based on fuzzy membership functions that specify the degree that a value in an event matches a value in a subscription. A-TOPSS does not consider schema approximation. Another work is S-TOPSS [24] which considers schema and value semantic matching. It proposes the use of agreed-upon ontologies and a system architecture that generates events other than the original ones by replacing concepts with taxonomic or ad-hoc related concepts. S-TOPSS provides a generic architecture but no concrete discussion or empirical validation has been provided. Generating new events out of the original ones has the disadvantage of overwhelming the system with large amount of events. The matching model in S-TOPSS is Boolean and scoring as a result of matching was not considered. FOMatch [29] proposes the use of fuzzy ontologies that all interacting parties agree upon. FOMatch is the closest to the work presented in this paper but it does not remove explicit semantic coupling from the system and does not free the user from using pre-defined vocabularies. Properties and values are handled indistinguishably and relatedness of terms is limited to a measure of combination of edge weights in a taxonomic and synonymy ontology.

Event Ranking

Previous work have considered ranking events according to range predicates [19], preference, diversity and freshness [9], probability of occurrence [28], fuzzy membership of attribute values [17] or focused on efficient ranking in sliding windows rather than the ranking functionality [25]. All of these works do not use semantic relatedness of events as a factor for ranking. FOMatch [29] considers scoring based on semantic matching and evaluation was conducted using thresholds, however a precision-recall tradeoff was not investigated.

8. FUTURE WORK

Approximate matching of events in general and approximate semantic matching in particular have impacts on further aspects of event-based systems. This section discusses implications on event enrichment and complex event processing.

8.1 Event Enrichment

Hinze et al. [14] states that "event enrichment calls for an understanding not only of the events but also for the external sources of information". Two challenges are recognized with event enrichment are: deciding which events to be enriched and using which data, and the maintenance of enrichment rules or queries. The proposed model of approximate semantic matching can be extended to enable a dynamic enrichment approach where external enrichment logic is removed. Semantic approximation can be used when selecting information for enrichment and when matching enriched events [13].

8.2 Complex Event Processing

Complex Event Processing (CEP) concerns deriving a complex event by combining base events using a specific set of event constructors such as disjunction, conjunction, sequence, etc. [18]. CEP typically is based on a crisp event pattern matching model where single events matching is decided in a Boolean manner and results proceed to the pattern matcher which tests different relationships of interest [18].

The proposed approximate semantic matching model suggests matching single events with scores rather than Boolean results. Those scores need to be propagated then in the pattern matching model and an appropriate model for concluding the score of complex events is needed. This means an uncertain matching model rather than an exact model. Two challenges associated with this are: defining the suitable semantics of scores and the appropriate model for pattern matching and ranking of complex events, and leveraging the pattern being matched for approximate semantic matching of single events. In other words, to affect the functions f_P, f_V and Top_m defined in the single event matching model.

9. CONCLUSIONS

Event-based systems are coupled, via event subscriptions and patterns, to the semantics of the underlying event schema and values. Approximate semantic matching of heterogeneous events has been discussed in this paper in order to address event semantic coupling. Event semantic of types, properties and values has been considered as a dimension of decoupling required to scale event-based systems out to high heterogeneous environments such as the sensor web. A general model has been proposed with a hybrid instantiation based on both thesauri and distributional semantics-based semantic similarity and relatedness measures. Experiments have been conducted on real-world events extracted from Wikipedia and Freebase. Results show that the proposed hybrid matcher outperforms matchers based on a single semantic similarity or relatedness measure.

10. ACKNOWLEDGMENTS

This work has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

11. REFERENCES

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

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web 4825*, (2007), 722-735.

[3] Aumueller, D., Do, H.-H., Massmann, S., and Rahm, E. Schema and ontology matching with COMA++. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM (2005), 906-908.

[4] Belkin, N.J. and Croft, W.B. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM 35*, 12 (1992), 29-38.

[5] Berners-Lee, T. Linked Data- Design Issues. 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[6] Botts, M., Percivall, G., Reed, C., and Davidson, J. OGC sensor web enablement: Overview and high level architecture. *GeoSensor networks*, (2008), 175-190.

[7] Budanitsky, A. and Hirst, G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics 32*, 1 (2006), 13-47.

[8] Curry, E., Hasan, S., Hassan, U. ul, Herstand, M., and O'Riain, S. An Entity-Centric Approach to Green Information Systems. *The 19th European Conference on Information Systems (ECIS)*, (2011).

[9] Drosou, M., Stefanidis, K., and Pitoura, E. Preference-aware publish/subscribe delivery with diversity. *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems*, (2009), 6:1--6:12.

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

[11] Freitas, A., Curry, E., J.G., O., and O'Riain, S. Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends. *IEEE Internet Computing Special Issue on Internet Scale Data 16*, 1 (2012), 24-33.

[12] Gabrilovich, E. and Markovitch, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on Artifical intelligence*, (2007), 1606-1611.

[13] Hasan, S., Curry, E., Banduk, M., and O'Riain, S. Toward Situation Awareness for the Semantic Sensor Web: Complex Event Processing with Dynamic Linked Data Enrichment. *the 4th International Workshop on Semantic Sensor Networks 2011 (SSN11)*, (2011), 60-72.

[14] Hinze, A., Sachs, K., and Buchmann, A. Event-based applications and enabling technologies. *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems*, ACM (2009), 1:1--1:15.

[15] Jiang, J.J. and Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of the International Conference on Re- search in Computational Linguistic*, (1998).

[16] Klyne, G. and Carroll, J.J. Resource Description Framework (RDF): Concepts and Abstract Syntax. 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

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

[18] Luckham, D.C. *The power of events: an introduction to complex event processing in distributed enterprise systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[19] Machanavajjhala, A., Vee, E., Garofalakis, M., and Shanmugasundaram, J. Scalable ranked publish/subscribe. *Proc. VLDB Endow. 1*, 1 (2008), 451-462.

[20] Manning, C.D., Raghavan, P., and Schutze, H. *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008.

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

[22] OECD. Machine-to-Machine Communications: Connecting Billions of Devices. *OECD Digital Economy Papers No. 192*, 2012.

[23] Patwardhan, S. and Pedersen, T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, (2006), 1-8.

[24] Petrovic, M., Burcea, I., and Jacobsen, H.-A. S-ToPSS: semantic Toronto publish/subscribe system. *Proceedings of the 29th international conference on Very large data bases - Volume 29*, VLDB Endowment (2003), 1101-1104.

[25] Pripužić, K., Žarko, I.P., and Aberer, K. Top-k/w publish/subscribe: finding k most relevant publications in sliding time window w. *Proceedings of the second international conference on Distributed event-based systems*, ACM (2008), 127-138.

[26] Prud'Hommeaux, E. and Seaborne, A. SPARQL query language for RDF. *W3C working draft 4*, January (2008).

[27] Rada, R., Mili, H., Bicknell, E., and Blettner, M. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on 19*, 1 (1989), 17-30.

[28] Wasserkrug, S., Gal, A., Etzion, O., and Turchin, Y. Complex event processing over uncertain data. *Proceedings of the second international conference on Distributed event-based systems*, (2008), 253-264.

[29] Zhang, W., Ma, J., and Ye, D. FOMatch: A Fuzzy Ontology-Based Semantic Matching Algorithm of Publish/Subscribe Systems. *Computational Intelligence for Modelling Control & Automation, 2008 International Conference on*, (2008), 111-117.