Towards Scalable Exploration of Diagnoses in an Ontology Stream

Freddy Lécué
IBM Dublin Research Center
Damastown Industrial Estate, Dublin, Ireland
{(firstname.lastname)@ie.ibm.com}

Abstract

Diagnosis, or the process of identifying the nature and cause of an anomaly in an ontology, has been largely studied by the Semantic Web community. In the context of ontology stream, diagnosis results are not captured by a unique fixed ontology but numerous time-evolving ontologies. Thus any anomaly can be diagnosed by a large number of different explanations depending on the version and evolution of the ontology. We address the problems of identifying, representing, exploiting and exploring the evolution of diagnoses representations. Our approach consists in a graph-based representation, which aims at (i) efficiently organizing and linking time-evolving diagnoses and (ii) being used for scalable exploration. The experiments have shown scalable diagnoses exploration in the context of real and live data from Dublin City.

Introduction

The semantic web (Berners-Lee, Hendler, and Lassila 2001) is considered to be the future of the current web. The semantics of information is represented using rich description languages e.g., OWL the *Web Ontology Language* (OWL Working Group 2009). OWL is underpinned by Description Logics (DL) (Baader and Nutt 2003) to define web ontologies. While automatic processing of semantics-augmented information can be achieved using state-of-the-art inference methods, these reasoning techniques are designed for static ontologies. However, information and knowledge are usually subject to change over time, even drastically in real world applications. Ontology versioning (Noy and Musen 2002), semantic sensors (Sheth 2010) are examples where knowledge is evolving on a time basis. Ontology stream (Huang and Stuckenschmidt 2005) i.e., time-based sequence of ontology, is one model to represent dynamic knowledge.

Diagnosis, or the process of identifying the nature and cause of an anomaly (a.k.a conflict or inconsistency) in an ontology, has been largely studied by the Semantic Web community. Existing works (Parsia, Sirin, and Kalyanpur 2005) applied and extended axioms pinpointing approaches (Baader, Peñaloza, and Suntisrivaraporn 2007) to derive how (fixed) ontology conflicts are propagated.

We focus on a time-based evolution of ontologies instead. In this context of ontology stream, any anomaly (as an instance) can be diagnosed by a large number of different explanations (as concept expressions) depending on the version of the ontology. In a traffic context, an anomaly is a congested road and its diagnoses could be be a road work, a music event or a road incident. These explanations can be derived at different time intervals, but no semantics is provided on their temporal evolution. Since the set of explanations is exponentially growing on a time basis, understanding anomalies and the profile of their diagnoses over time is a complex task. Appropriate knowledge models are then required to explore diagnoses results in a scalable way.

We address the problems of identifying, representing, exploiting and exploring the evolution of diagnoses representations using their implicit semantic relations. Towards this issue we present the directed diagnoses graph (DRG), which aims at efficiently organizing and linking time-evolving diagnoses. The links are captured by (subsumption-based) semantic relations between diagnoses while their time-based evolution and changes are interpreted by constructive reasoning abduction (Noia et al. 2003). The DRG is then used for scalable exploration, through subsumption and abduction relations, of time-based evolving diagnoses. The experiments have shown scalable diagnoses exploration in the context of real and live data from Dublin City in Ireland.

In the following we review the logic adopted together with ontology stream and diagnosis. Then we present how diagnosis results are linked in a compact DRG. The next section presents how this graph is exploited to explore diagnoses over time. Then we report experiments in the context of real and live data from Dublin City. Finally, we briefly comment on related work and draw some conclusions.

Background

We focus on DL as formal knowledge representation language to define ontologies streams and the underlying background knowledge. We review (i) DL basics of $\mathcal{EL}^{++}$, (ii) ontology stream, (iii) anomaly, its (iv) diagnosis problem.

$\mathcal{EL}^{++}$ Description Logics

We illustrated our work with DL $\mathcal{EL}^{++}$ where satisfiability and subsumption are decidable. The selection of this DL fragment has been guided by the expressivity which was required to model semantics of data in our domain e.g., transportation traffic, event and road works data. The DL $\mathcal{EL}^{++}$

(Baader, Brandt, and Lutz 2005) is the logic underpinning OWL 2 EL and the basis of many more expressive DL. Adaptations of our approach to more expressive DLs could be possible to some extent (e.g., DL with cardinality restrictions, concept union, negation) but would impact decidability and complexity (cf. comments in Validation Section).

A signature Σ, defined by $(\mathcal{N}_C, \mathcal{N}_R, \mathcal{N}_I)$, consists of 3 disjoint sets of (i) atomic concepts $\mathcal{N}_C$, (ii) atomic roles $\mathcal{N}_R$, and (iii) individuals $\mathcal{N}_I$. Given a signature, the top concept $\top$, the bottom concept $\bot$, an atomic concept A, an individual a, an atomic role r, $\mathcal{EL}^{++}$ concept expressions C and D can be composed with constructs: $\top \mid \bot \mid A \mid C \sqcap D \mid \exists r.C \mid \{a\}$. We denote by $\mathcal{E}_C$ this set of concept expressions. We slightly abuse the notion of atomic concepts to include $\top$, $\bot$, nominals (Horrocks and Sattler 2001) i.e., individuals appearing in concept definitions of form $\{a\}$. The particular DL-based ontology $\mathcal{O} \doteq< \mathcal{T}, \mathcal{A} >$, is composed of a static TBox $\mathcal{T}$ (concept, role axioms), and ABox $\mathcal{A}$.

Example 1. ***($\mathcal{EL}^{++}$ DL Concept)***
According to the TBox $\mathcal{T}$ in Figure 1, a $CongestedRoad$ is "a road with at least a bus which is heavily congested" (2).

$$\exists from.Area \sqcap \exists to.Area \sqcap \exists travel.Bus \sqsubseteq Road \sqcap \exists with.Bus \quad (1)$$
$$Road \sqcap \exists with.(Bus \sqcap \exists congested.Heavy) \sqsubseteq CongestedRoad \quad (2)$$
$$Road \sqcap \exists with.(Bus \sqcap \exists congested.Light) \sqsubseteq FreeRoad \quad (3)$$
$$CongestedRoad \sqcap FreeRoad \sqsubseteq \bot \quad (4)$$
$$(Road \sqcap \exists junction.\{r_3\} \sqcap \exists venue.\{TheO2\})(r_1) \quad (5)$$
$$Road(r_3) \quad (6)$$

Figure 1: Static $\mathcal{EL}^{++}$ TBox $\mathcal{T}$ (1-4) and ABox $\mathcal{A}$ (5-6).

$\mathcal{EL}^{++}$ supports General Concept Inclusion axioms (GCIs) e.g. $C \sqsubseteq D$ with C is subsumee and D subsumer and role inclusion axioms (RIs, e.g., $r \sqsubseteq s, r_1 \circ \cdots \circ r_n \sqsubseteq s$). An ABox is a set of concept assertion axioms e.g., $C(a)$, role assertion axioms e.g., $R(a, b)$, and individual in/equality axioms e.g., $a \neq b$ or $a = b$. We assume that $\mathcal{EL}^{++}$ TBox is normalized, and all subsumption closures are pre-computed (Baader, Brandt, and Lutz 2005). We use the term background knowledge to refer to such TBoxes.

Ontology Stream

We represent knowledge evolution through a dynamic, evolutive version of ontologies (Huang and Stuckenschmidt 2005) i.e., ontology stream (Definition 1).

Definition 1. ***(Ontology Stream)***
An ontology stream $\mathcal{O}_1^n$ from point of time 1 to time n is a sequence of ABox axioms $(\mathcal{O}_1^n(1), \mathcal{O}_1^n(2), \cdots, \mathcal{O}_1^n(n))$ with respect to a static and fixed Tbox $\mathcal{T}$ where $n \geq 1$.

$\mathcal{O}_1^n(i)$ is a snapshot of an ontology stream (stream for short) $\mathcal{O}_1^n$ at time i, referring to ABox axioms with respect to $\mathcal{T}$.

Example 2. ***(Ontology Stream)***
Figure 2 illustrates snapshots of $\mathcal{O}_1^7$ through ABox axioms. $bus7$ is in a light and heavy traffic in $\mathcal{O}_1^7(6)$ and $\mathcal{O}_1^7(7)$.

Anomaly in Ontology Stream

An anomaly, capturing an abnormal situation in a stream, is defined as any instance of two disjoint concept expressions, one at point of time i; and the other at point of time $i+1$.

Definition 2. ***(Anomaly in Ontology Stream)***
Let $\mathcal{O}_1^n$ be a stream; $\mathcal{T}$ be a set of Tbox axioms; a be an individual in $\mathcal{N}_I$. The individual a is an anomaly in $\mathcal{O}_1^n$ at time $i \in [1, n)$ iff $\exists B, C \in \mathcal{E}_C \backslash \{\bot\}$ such that:

$$\mathcal{T} \cup \mathcal{O}_1^n(i) \models B(a) \quad (7) \qquad \mathcal{T} \cup \mathcal{O}_1^n(i+1) \models C(a) \quad (8)$$
$$\mathcal{T} \models B \sqcap C \sqsubseteq \bot \quad (9)$$

Example 3. ***(Anomaly in Ontology Stream)***
Road r_1 is defined as an anomaly at time 6 of stream $\mathcal{O}_1^7$. Indeed it is straightforward to derive that individual r_1 is a $FreeRoad$ using (1)*,* (3)*, (10-11) and then a $CongestedRoad$ using (1-2), (13-14) in respectively $\mathcal{O}_1^7(6)$ and $\mathcal{O}_1^7(7)$, where both descriptions are disjoint* (4) *in $\mathcal{T}$.*

The anomalies are derived by capturing dynamic knowledge from the stream (Figure 2), and then interpreted using the background knowledge (Figure 1). Updating, adding and removing anomalies that need to be captured is straightforward, as it mainly requires axioms with the semantics of (4).

$$\mathcal{O}_1^7(6) : (Bus \sqcap \exists congested.Light)(bus7) \quad (10)$$
$$: (\exists from.\{x\} \sqcap \exists to.\{y\} \sqcap \exists travel.\{bus7\})(r_1) \quad (11)$$
$$: (\exists liveOnStage.\{U2\})(theO2) \quad (12)$$
$$\mathcal{O}_1^7(7) : (Bus \sqcap \exists congested.Heavy)(bus7) \quad (13)$$
$$: \exists from.\{x\} \sqcap \exists to.\{y\} \sqcap \exists travel.\{bus7\}(r_1) \quad (14)$$
$$: (\exists roadWorks.Resurface)(r_3) \quad (15)$$

Figure 2: Stream Snapshots of $\mathcal{O}_1^7$: $\mathcal{O}_1^7(6)$ and $\mathcal{O}_1^7(7)$.

Semantics-based Diagnosis

Adapted in ontology stream, Definition 3 (Lécué 2012) captures a semantics-based diagnosis problem where stream-based anomalies are captured by Definition 2.

Definition 3. ***(Semantic Diagnosis Problem - SDP)***
Let $\mathcal{T}$ be TBox axioms; $\mathcal{O}_1^n$ be a stream; a be an anomaly of $\mathcal{O}_1^n$ at time $n-1$. A diagnosis problem $SDP\langle \mathcal{T}, \mathcal{O}_1^n, a, k \rangle$ consists in finding all expressions E in $\mathcal{E}_C \backslash \{\bot\}$ under k-window $[n-k, n-1]$, with $k \in [1, n-1]$ such that:

$$\mathcal{O}_{n-k+1}^n \cup \mathcal{T} \models E(a) \quad (16) \qquad \mathcal{O}_{n-k+1}^n \cup \mathcal{T} \models C(a) \quad (17)$$
$$\mathcal{T} \not\models E \sqcap C \sqsubseteq \bot \quad (18)$$

Contrary to (Horridge, Parsia, and Sattler 2008) where the diagnose is identified as a set of axioms, our diagnosis task consists in finding all non-disjoint concept expressions describing anomaly a at different time intervals (from a k-window in $\mathcal{O}_1^n$ (16)) using dynamic ABoxes in $\mathcal{O}_{n-k+1}^n$ and static background knowledge $\mathcal{T}$. Following (Lécué 2012) a diagnose can be interpreted as a set of various representations of a over time. These non-disjoint representations are potential explanations of the anomaly. We interpret them as diagnoses since they provide alternative ways (or expressions) of representing the anomaly a. The axioms from the evolving knowledge are crucial to find E. (18) discards expressions which may cause a to be an anomaly (Definition 2). The most specific expression under $\sqsubseteq$ is returned for each fixed k. Deciding if a concept expression belongs to solutions of a SDP is PTIME-hard with respect to polynomial k and n while constructing a solution is PSPACE-hard in $\mathcal{EL}^{++}$ (Lécué 2012).

Example 4. ***(Semantic Diagnosis Problem -*** SDP***)***
Following Example 3, we have r_1 *as an anomaly in* $\mathcal{O}_1^7(7)$:

$$\mathcal{O}_1^7(7) \cup \mathcal{T} \models CongestedRoad(r_1) \quad (19)$$

Let $\langle \mathcal{T}, \mathcal{O}_1^7, r_1, k \rangle$*, with* $k = 2$ *be a* SDP*. The explanations (as concept expressions)* E *of* r_1 *to be congested are different whether we consider* $\mathcal{O}_1^7(6)$ *in* (20) *using axioms* (5)*,* (12)*; or* $\mathcal{O}_1^7(7)$ *in* (21) *using* (5)*,* (6)*,* (15)*; or both in* (22)*.*

$$\mathcal{O}_1^7(6) \cup \mathcal{T} \models (\exists venue.(\exists liveOnStage.\{U2\}))(r_1) \quad (20)$$

$$\mathcal{O}_1^7(7) \cup \mathcal{T} \models (\exists junction. (\exists roadWorks.Resurface))(r_1) \quad (21)$$

$$\cup_{i=6}^{7} \mathcal{O}_1^7(i) \cup \mathcal{T} \models (\exists venue.(\exists liveOnStage.\{U2\}) \sqcap \exists junction.(\exists roadWorks. Resurface))(r_1) \quad (22)$$

Linking Diagnoses in an Ontology Stream

The number of diagnoses could be very large depending on the k-window given in a SDP (Definition 3). Indeed a large number of expressions can be interpreted as an explanation of any anomaly. Apart from their temporal connection, it is not straightforward to derive how explanations are properly linked. We provide a compact representation to efficiently organize diagnoses in our context (Figure 3).

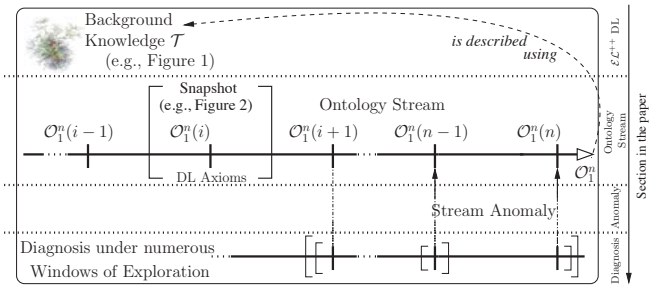

Figure 3: Diagnosing Anomalies in Ontology Stream.

Directed Diagnoses Graph

In case of an anomaly at point of time $n-1$ and n in $\mathcal{O}_1^n$, a number of $n \times (n+1)/2$ windows of observations and underlying SDPs (Definition 3) are required to be solved. For instance $\{1\}$, $\{2\}$, $\{3\}$, $[1,2]$, $[2,3]$ and $[1,3]$ are all the windows for $n = 3$. Definition 4 organizes the diagnoses along a directed graph where their links are established through subsumption relationships. This ensures to easily track specialization and generalization-based evolution of diagnoses.

Definition 4. ***(Directed Diagnoses Graph -*** DRG***)***
Let (i) $\mathcal{O}_1^n$ *be a stream; (ii)* $\mathcal{T}$ *be TBox axioms; (iii)* a *be an anomaly at point of time* $n-1$ *of* $\mathcal{O}_1^n$*. A* DRG $D = \langle V, R \rangle$ *of* $SDP\langle \mathcal{T}, \mathcal{O}_1^n, a, k \rangle$*,* $k \in [1,n]$*, is a directed graph where:*

- *its vertices set* V *is defined by:*

$$\{E_i^j \mid E_i^j \text{ is a result of } SDP\langle \mathcal{T}, \mathcal{O}_i^j, a, j-i+1 \rangle, \forall i,j \in [1,n] \text{ with } i \leq j\} \quad (23)$$

- *its arcs* R *is defined by a set of* (E_g^h, E_i^j) *iff:*

$$\mathcal{O}_g^h \cup \mathcal{O}_i^j \cup \mathcal{T} \models E_g^h \sqsubseteq E_i^j \quad (24)$$

$$\nexists [p,q] \in [1,n] \mid \mathcal{O}_g^h \cup \mathcal{O}_p^q \cup \mathcal{O}_i^j \cup \mathcal{T} \models E_g^h \sqsubseteq E_p^q \sqsubseteq E_i^j \quad (25)$$

with E_g^h, E_p^q, E_i^j *are concept expressions;* $[g,h]$*,* $[p,q]$*,* $[i,j]$ *are windows in* $[1,n]$ *such that* $[p,q] \neq [g,h]$*,* $[p,q] \neq [i,j]$*.*

For the sake of clarity we assume E_i^j to be the most specific expression of SDP in (23). This can be generalized to sets by considering each of its expression as a vertex in D.

A DRG organizes all $n \times (n+1)/2$ possible results of all $n \times (n+1)/2$ SDPs (Definition 3), all computed under a window of $k \in [1,n]$ snapshots. The results, captured by V, are linked only in case of subsumption (24). All arcs, which can be inferred through the transitive property of DL subsumption, are not captured by a DRG (25), thus minimizing the amount of arcs explicitly stored. Indeed, omitting (25) would reach to DRGs with up to $n^2 \times (n^2-1)/4$ arcs. Alternatively, the number of arcs of DRGs with (25) is bounded by $n \times (n+1) - 2$ i.e., a quadratic number instead.

Example 5. ***(Directed Diagnoses Graph)***
Let $\mathcal{O}_1^7(5)$ *be a snapshot of* $\mathcal{O}_1^7$ *with same axioms as in* $\mathcal{O}_1^7(6)$*. Suppose the (most specific) results of each* $SDP\langle \mathcal{T}, \mathcal{O}_5^7, r_1, k \rangle$*, with* $k \leq 3$*, as follows. Explanations (i)* E_5^5*,* E_5^6*,* E_6^6 *are described by concept expression of* (20)*, (ii)* E_7^7 *is described by expression of* (21)*, (iii)* E_5^7*,* E_6^7 *are described by expressions of* (22)*. Its* DRG *(Figure 4) reflects how diagnoses are evolving e.g.,* $\mathcal{O}_6^7 \cup \mathcal{T} \models E_6^7 \sqsubseteq E_7^7$*. By transitivity of subsumption, we have* $\mathcal{O}_5^6 \cup \mathcal{T} \models E_5^5 \sqsubseteq E_5^6$*.*

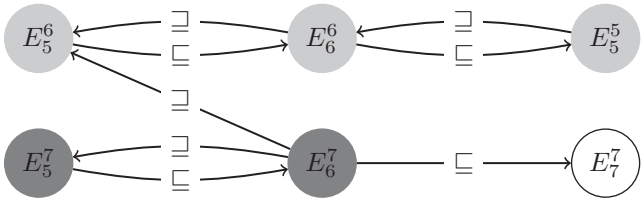

Figure 4: Illustration of a Directed Diagnoses Graph.

Arcs Contraction

Similar diagnoses could be derived from different windows. Such similarities (captured by equivalence of expressions) are due to knowledge remaining invariant over time e.g., windows $[5,7]$ (i.e., E_5^7) and $[6,7]$ (i.e., E_6^7) in Example 5 and Figure 4. Definition 5 compiles this knowledge by contracting arcs of DL equivalent results.

Definition 5. ***(Arcs-Contraction)***
Let $D : \langle V, R \rangle$ *be a* DRG *containing arcs* $\{(u,v),(v,u)\}$ *with* $u \neq v$*. Let* f *be a function which maps every vertex in* $V \backslash \{u,v\}$ *to itself, and otherwise, maps it to a new vertex* w*. The contraction of* $\{(u,v),(v,u)\}$ *results in a* DRG $D' = \langle V', R' \rangle$*, where* $V' = (V \backslash \{u,v\}) \cup \{w\}$*,* $R' = R \backslash \{(u,v),(v,u)\}$*, and for each* $x \in V$*,* $x' = f(x) \in V'$ *is incident to an arc* $r' \in R'$ *iff, the corresponding arc,* $r \in R$ *is incident to* x *in* D*.*

An arcs-contraction is a task which removes arcs connecting two DL equivalent diagnoses from a DRG while simultaneously merging together the two vertices it previously connected. Contracting arcs does not modify knowledge encoded in the initial DRG, it actually re-arranges it for (i)

minimizing its size, thus minimizing its exploration time, (iii) identifying windows with similar results. A contraction of a DRG, called *arcs-contracted* DRG, is the result of a sequence of arcs-contractions where vertices are labelled by windows and described by their diagnoses.

Example 6. ***(Arcs-Contraction)***
All arcs of DRG in Example 5 which connect vertices of DL equivalent diagnoses (vertices with similar grey level in Figure 4) have been merged in Figure 5.

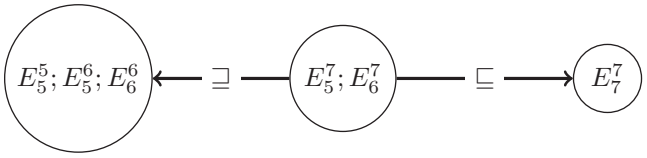

Figure 5: Arcs-Contraction of DRG in Example 5.

Organizing results from SDPs $\langle \mathcal{T}, \mathcal{O}_1^n, A, k\rangle$ consists in (i) fixing the window $[1, n]$, which bounds the search space of diagnoses, (ii) evaluating DRGs (Definition 4), which connects diagnoses by subsumption, (iii) computing its arcs-contraction (Definition 5). The complexity of this organization is $\Theta(n^2)$ where n is the number of results of the SDP.

Following the graph-based representation of diagnoses evolution over time, merging DRGs, evaluating the shortest path between two diagnoses can be achieved by adapting state-of-the-art graph theory techniques (Gross and Yellen 2005). E.g., the fusion of DRGs consists in incrementally connecting vertices based on their subsumption relationship.

Diagnoses Exploration

We present how abduction, as a way to reveal difference-based evolution of diagnoses over time, can be efficiently computed in a DRG. Then we present how such an augmented DRG is explored for detecting relevant diagnoses.

Interpreting Changes in Diagnoses

Even if a DRG links its diagnoses by subsumption, this relation fails in capturing qualitatively each diagnose evolution. Towards this issue, we adapt constructive DL reasoning abduction (Noia et al. 2003) between diagnoses E_g^h and E_i^j in Definition 6 and extend the DRG definition.

Definition 6. ***(Concept Abduction Problem -*** CAP***)***
Let (i) $\mathcal{O}_1^n$ be a stream; (ii) $\mathcal{O}_g^h$, $\mathcal{O}_i^j \subseteq \mathcal{O}_1^n$ where $g, h, i, j \in [1, n]$ with $g \leq h$, $i \leq j$; (iii) E_g^h, E_i^j be two concept expressions in $\mathcal{E}_C$; (iv) $\mathcal{T}$ be TBox axioms. A CAP $E_g^h \backslash E_i^j$ consists in finding an expression $X \in \mathcal{E}_C$ such that:

$$E_i^j \sqcap X \sqsubseteq E_g^h \quad (26) \qquad \mathcal{T} \not\models E_i^j \sqcap X \sqsubseteq \bot \quad (27)$$

$$\nexists X' \in \mathcal{E}_C \mid E_i^j \sqcap X' \sqsubseteq E_g^h \text{ and } X \sqsubseteq X' \quad (28)$$

with respect to $\mathcal{O}_g^h \cup \mathcal{O}_i^j \cup \mathcal{T}$.

Definition 6, through abduction, captures what is over-specified by diagnose E_g^h with respect to E_i^j in $\mathcal{O}_g^h \cup \mathcal{O}_i^j \cup \mathcal{T}$. The most general expression (28) is computed to obtain $E_i^j \sqcap X$ and E_g^h as "*close*" as possible under subsumption. A trivial solution of a CAP $E_g^h \backslash E_i^j$ where $E_i^j \sqsubseteq E_g^h$ is $\top$ i.e., no relevant information is returned as no expression is over-specified by E_g^h with respect to E_i^j.

Example 7. ***(Concept Abduction Problem -*** CAP***)***
Let E_7^7, E_5^7 be $\mathcal{EL}^{++}$ expressions respectively defined by (21)*,* (22) *in $\mathcal{O}_1^n$. From E_5^7 to E_7^7, the diagnose denoted by $E_5^7 \backslash E_7^7$ in* (29)*, is missed i.e., some representations are not returned as part of the diagnose in E_7^7 while they are in E_5^7.*

$$E_5^7 \backslash E_7^7 \doteq \exists venue.(\exists liveOnStage.\{U2\}) \quad (29)$$

Definition 7 applies CAP to all arcs A of DRGs, so diagnoses changes among adjacent results can be interpreted.

Definition 7. ***(Abduction-Annotated*** DRG***)***
Let $D = \langle V, R\rangle$ be a DRG. The abduction-annotated DRG of D is defined by $\langle V, R, f\rangle$ where f is a function which maps every arc $(E_g^h, E_i^j) \in R$ to $E_g^h \backslash E_i^j$.

Example 8. ***(Abduction-Annotated*** DRG***)***
$D' = \langle V, R, f\rangle$ is the abduction-annotated DRG of D illustrated in Example 6 and Figure 5 where f maps all $\mathcal{EL}^{++}$ equivalent arcs (E_5^7, E_7^7) to (29) *and (E_6^7, E_6^6) to* (21)*.*

Abduction is required for evaluating the evolution of diagnoses between not only adjacent results in E, by also between any diagnoses in V of a DRG $\langle V, R, f\rangle$. However, considering all pair of diagnoses would reach to a number of $n^2 \times (n^2 - 1)/4$ CAPs in $PSPACE$ (Noia et al. 2003) to be solved. This is problematic for streams with a large number of axioms. We override the direct computation of PSPACE CAPs by applying results of Theorem 1.

Theorem 1. ***(Abduction Reconstruction)***
Let (i) $\mathcal{T}$ be TBox axioms, (ii) A, B, C be concept expressions $\mathcal{E}_C$ satisfiable in $\mathcal{T}$. Table 1 describes rules which are used to evaluate abduction between vertices (expressions) which are linked through DRG arcs (Figure 6(a-b-c)).

	Rule ID, Description	Figure
R_1	If $A \sqsubseteq B,\ B \sqsubseteq C$ then $(A\backslash B) \sqcap (B\backslash C) \sqsubseteq (A\backslash C)$	6(a)
R_2	If $A \sqsubseteq B,\ C \sqsubseteq B$ then $(A\backslash B) \sqsubseteq (A\backslash C)$	6(b)
R_3	If $B \sqsubseteq A,\ B \sqsubseteq C$ then $(B\backslash C) \sqsubseteq (A\backslash C)$	6(c)

Table 1: Abduction Reconstruction in $\mathcal{T}$.

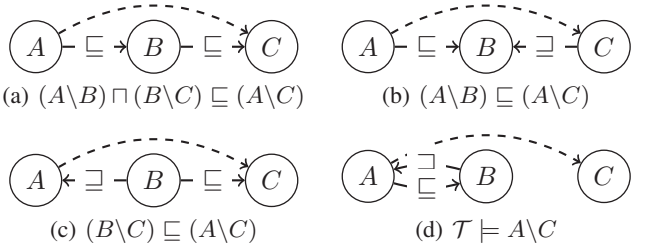

Figure 6:Abduction Reconstruction (DRG Illustration).

Proof. (Overview) (R_1) By Definition of $A\backslash B$ and $B\backslash C$, $\exists X, Y \in \mathcal{E}_C \mid B \sqcap X \sqsubseteq A$ and $C \sqcap Y \sqsubseteq B$. Thus, $(C \sqcap Y) \sqcap X \sqsubseteq A$. So $\exists Z \doteq Y \sqcap X \mid C \sqcap Z \sqsubseteq A$ hence Z as a solution of $A\backslash C$. (R_2) By definition of $A\backslash B$, $\exists X \in \mathcal{E}_C \mid B \sqcap X \sqsubseteq A$. As $C \sqsubseteq B$ in $\mathcal{T}$ then $C \sqcap X \sqsubseteq A$. So X is a solution of $A\backslash C$. (R_3) By definition of $B\backslash C$, $\exists X \in \mathcal{E}_C \mid C \sqcap X \sqsubseteq B$. As $B \sqsubseteq A$ in $\mathcal{T}$ then $C \sqcap X \sqsubseteq A$. X is a solution of $A\backslash C$. $\square$

Theorem 1 is proved by exploiting properties of subsumption and abduction. It could be used for reconstructing some solutions of CAPs $A\backslash C$ (cases (a-b-c) in Figure 6) in constant time and space e.g., context with known CAPs $A\backslash B$ of adjacent diagnoses A, B. However some cases e.g., (d) requires the CAP to be constructed directly. It is straightforward to generalize Theorem 1 to more complex paths between A and C by combining the three different cases.

Abduction-Guided Diagnoses Exploration

Algorithm 1 returns a list of *high-level* diagnoses (line 2) of a SDP (line 1), where only the most specific (line 10) and general results (line 11) under subsumption have been extracted from its DRG (line 7). The computation of their abduction (Definition 6), which requires flexible and fast reconstruction from existing ones (Theorem 1), is established for evaluating difference between comparable results (line 15). Since abduction is not necessarily unique, the expression with the minimal size (Küsters 2001) is returned.

Algorithm 1: $HighLevelDiagnosesViewer\langle \mathcal{T}, \mathcal{O}_1^n, A\rangle$.

1 **Input**: (i) $\mathcal{T}$ be TBox axioms; (ii) Stream $\mathcal{O}_1^n$; (iii) Anomaly A in $\mathcal{E}_C$, (iv) $k \in [1, n]$, (v) $SDP\langle \mathcal{T}, \mathcal{O}_1^n, A, k\rangle$ S.
2 **Result**:$\{(F, G, X) \in \mathcal{E}_C^3 \mid F, G$: results of S; X: abduction$\}$.
3 **begin**
4 $s \leftarrow \emptyset$; $g \leftarrow \emptyset$; *//Initializing most specific/general results.*
5 $sol \leftarrow \emptyset$; *// Initializing the result set.*
6 *// Abduction-annotated DRG of S: Definitions 4, 5, 7*
7 $\langle V, R, f\rangle \leftarrow$ Abduction-annotated DRG of SDP S;
8 *// Most specific, general results of S w.r.t. $\sqsubseteq$ are captured*
9 **foreach** $V_i \in V$ **do**
10 **if** $\nexists V_j \in V \mid \mathcal{O}_1^n \cup \mathcal{T} \models V_j \sqsubseteq V_i$ **then** $s \leftarrow s \cup V_i$;;
11 **if** $\nexists V_j \in V \mid \mathcal{O}_1^n \cup \mathcal{T} \models V_i \sqsubseteq V_j$ **then** $g \leftarrow g \cup V_i$;;
12 *// High-level view and classification of results*
13 **foreach** $C_1 \in s$ *and* $C_2 \in g$ **do**
14 *// Subsumption between a most specific/general result*
15 **if** $\mathcal{O}_1^n \cup \mathcal{T} \models C_1 \sqsubseteq C_2$ **then**
16 *// Association of results, valued by abduction*
17 $sol \leftarrow sol \cup (C_1, C_2, C_1\backslash C_2)$;
18 **return** sol;

The complexity of Algorithm 1 is strongly correlated to the structure of the underlying abduction-annotated DRG i.e., $|E|$ PSPACE-hard abduction problems to be solved in the worst case. If all arcs are abduction-annotated, its complexity is in PTIME using properties of Theorem 1 i.e., $\Theta(n^4)$ subsumptions tests (line 9), PSPACE otherwise.

The exploration of diagnoses, which could be performed by end-users, consists in (i) accessing the output list of Algorithm 1, which is more convenient than the complete list of candidate results, (ii) exploring, iterating through specialization/generalization of results via adjacent vertices using initially the previous list, (iii) interpreting the diagnoses evolution (through abduction) during the exploration. Thus, abduction is used as a guide through the exploration. As results are organized under subsumption in a DRG, the exploration of more specific or general diagnoses is straightforward. For the same reason, getting their abduction is important for tracking specialization/generalization changes. End-user could validate, reject results and finalize the exploration at any point. In case of rejected results, any more specific diagnose is automatically removed from the search space.

Example 9. ***(Abduction-Guided Diagnoses Exploration)*** *The diagnoses are explored using the abduction-annotated DRG (Example 8 - Figure 5 with abduction on arcs). Following Algorithm 1 the most specific and general diagnoses are retrieved together with their abduction. The exploration can start from the latter elements and be iterated using the specialization/generalization features of the DRG. E.g.,* (29) *gives the semantics of moving from E_5^7 to E_7^7.*

Validation

This section (i) reports the degree of compactness of diagnoses using our graph-based representation, and (ii) shows the scalability of its diagnoses exploration. We diagnose anomalies (Definition 2) using real live streams on different windows. CEL (Baader, Lutz, and Suntisrivaraporn 2006) is used to check satisfiability, subsumption, while MAMAStng (Noia, Sciascio, and Donini 2007) constructs abduction of diagnoses. Experiments were run on a server of 4 Intel(R) Xeon(R) X5650, 2.67GHz cores, 6GB RAM.

Context: Dublin City

• **Dynamic knowledge:** XML stream data from Dublin Bus has been enriched with $\mathcal{EL}^{++}$ representations. GPS location, congestion status of 1000 buses (updated every 20 second) were axiomatized in a streaming ABox e.g., (10-11) through 12 RDF triples. All anomalies, described following Example 3, are captured from the Bus stream. We have (on average) 3280 distinct anomalies over a period of 180 minutes. City events e.g., concert were captured through events search engines *Eventful*, *EventBrite* where an average of 187 events. Each is described through 16 RDF triples. An average of 51 road works, car incidents a day have also been enriched through 16 RDF triples each.
• **Static knowledge:** A core static ontology of 55 concepts with 19 properties (17 concepts subsume the 38 remaining ones with a maximal depth of 3) has been considered.

Experiments

• **Diagnoses Compactness:** Figure 7 illustrates the (i) factor of compactness between a DRG and its compact representation $CDRG$. On average a $CDRG$ compacts the number of vertices and arcs with a factor of 10 and 35 respectively. The factor is exponential with the size of the stream window.

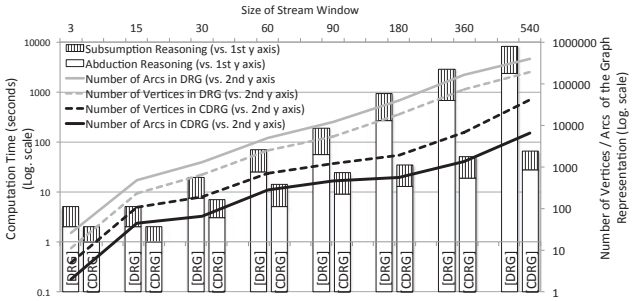

Figure 7: DRG Computation Time over Stream Windows.

We sketch the computation time required for (i) constructing their arcs (w.r.t. subsumption), (ii) elaborating the difference (w.r.t. abduction) between their vertices. The more compact the DRG the less its computation. The size of stream window does not impact negatively subsumption and abduction in a $CDRG$ as much as in a DRG. The overall computation performance is mainly impacted by the expressivity of the DL used, e.g., the computation of subsumption, abduction with more expressive DLs would have strongly altered (i) scalability (Algorithm 1) and even (ii) decidability in some cases (e.g., open issue for abduction in $\mathcal{SROIQ}$).

• **Diagnoses Exploration:** Figure 8 illustrates the benefits of exploring diagnoses using our approach. We evaluate the (average) number of iterations which is required by 50 non technical users for reaching a list of diagnoses using their links given a random initial diagnose. All users have been introduced the concepts of specialization, abduction, which are the basis of diagnoses (i) links, (ii) exploration. The diagnoses list, given by transportation experts, are the most representative explanations of congestion. The window of exploration of the dynamic knowledge is experimented from a range of 1 min (3 snapshots) to 180 min (540 snapshots). The iteration process is achieved using (a) the annotated DRG, (b) a naive structure where all diagnoses are not compressed or organized. Despite an initial quadratic number of diagnoses, only (average) 6 iterations are required to reach targeted diagnoses in a window of 30 snapshots. Users are able to make use of abduction results to easily navigate through their specialization. The maximum number of iterations is 51, which is not high given the search space, but too much for users. Experiments (not reported here) have shown that most of relevant diagnoses (85%) are within a window of 30 snapshots, which makes the approach scalable.

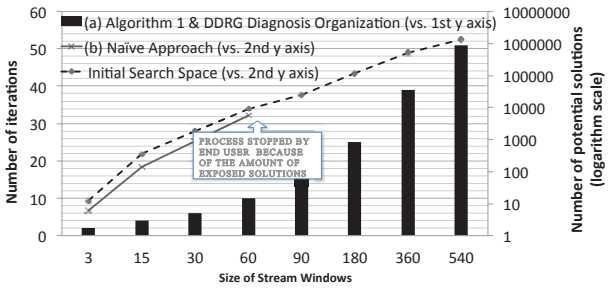

Figure 8: Diagnoses Exploration through DRG Iterations.

Discussion

We illustrated the approach in a road traffic context where anomalies are congested buses in Dublin City. It strongly helps in scaling up the diagnosis computation, specifically when the diagnosis is performed on multiple windows, and when a large amount of data and updates is considered. Since the Dublin Buses information is updated on a regular basis, its diagnoses are also updated. Maintaining a (semantic) coherent and scalable structure which links diagnoses evolution over time is crucial to (i) understand their evolution, and (ii) easily iterate on their links. The approach is general, and can be applied to any other context where sensors and data streams are involved. The (i) anomalies are any abnormal behavior or value in a stream i.e., Definition 2, Example 3, (ii) diagnoses are retrieved within a dynamic window of time (Definition 3). E.g., we could imagine the diagnosis of any fault or malfunctioning device.

All stream data is transformed in $\mathcal{EL}^{++}$, but only the diagnoses and the compact representation of their abduction are captured in the DRG. Both the (i) computation of diagnoses, (ii) size of the DRG are impacted by the amount of heterogeneous data. It is crucial to target relevant stream data in a diagnosis problem. Capturing and representing irrelevant data sources may end up to noisy diagnoses.

Related Work

Conflicts diagnosis in ontologies is widely studied (Parsia, Sirin, and Kalyanpur 2005). The objective is understanding how changes in an ontology may result in conflicts in the knowledge base. Various techniques have been introduced to infer which axioms are responsible of propagating errors and injecting conflicts. Existing approaches have been largely influenced by axioms pinpointing (Baader, Peñaloza, and Suntisrivaraporn 2007) and subsumption explanation (McGuinness and Borgida 1995), which consists in explaining (i) how knowledge is articulated in an ontology, (ii) what are the dependencies between concepts, (iii) which axioms are required to reach an entailment. Instead, we focus on a time-based evolution of ontologies and address the problem of organizing diagnoses in an evolving knowledge. In addition our diagnosis task consists in identifying expressions of explanation at various time interval rather than axioms.

(Lécué 2012) introduced diagnosis in an ontology stream, but assumed a pre-determined fixed time window where all causes are identifiable. This approach consists in extracting a common description from this unique window of exploration. This assumption cannot be verified in most of real-world problems. Indeed, how to determine whether the causes of a congestion can be retrieved within the last 5 or 30 minutes? From a different angle, (Fanizzi, d'Amato, and Esposito 2008) address changes of concept description and novelty detection using unsupervised machine learning techniques, which is based on clustering of new individuals. In another context, model-based diagnosis (Torta and Torasso 2003) identifies anomalies and their explanations by interpreting the behavior of a system model. All the latter techniques do not capture temporal evolution of faults and their diagnoses, limiting the interpretation of any time-aware evolution. As a complimentary approach we compact and annotate diagnoses evolution using a graph-based structure.

Conclusion and Future Work

We studied diagnosis i.e., identification of the nature and cause of an anomaly. We addressed the problem of identifying, representing, exploiting, exploring the evolution of diagnoses representations in a context of ontology stream. Towards this issue we present the directed diagnoses graph (DRG), which aims at (i) efficiently organizing, linking time-evolving diagnoses, (ii) being used for scalable exploration, through subsumption and abduction relations.

In future work, we will apply advanced semantic matching functions (Li and Horrocks 2003) for linking diagnoses.

Acknowledgments

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement ID 318201 (SIMPLI-CITY).

References

Baader, F., and Nutt, W. 2003. In *The Description Logic Handbook: Theory, Implementation, and Applications*.

Baader, F.; Brandt, S.; and Lutz, C. 2005. Pushing the el envelope. In *IJCAI*, 364–369.

Baader, F.; Lutz, C.; and Suntisrivaraporn, B. 2006. Cel - a polynomial-time reasoner for life science ontologies. In *IJCAR*, 287–291.

Baader, F.; Peñaloza, R.; and Suntisrivaraporn, B. 2007. Pinpointing in the description logic el^+. In *KI*, 52–67.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web. *Scientific American* 284(5):34–43.

Fanizzi, N.; d'Amato, C.; and Esposito, F. 2008. Conceptual clustering and its application to concept drift and novelty detection. In *ESWC*, 318–332.

Gross, J. L., and Yellen, J. 2005. *Graph Theory and Its Applications*.

Horridge, M.; Parsia, B.; and Sattler, U. 2008. Laconic and precise justifications in owl. In *International Semantic Web Conference*, 323–338.

Horrocks, I., and Sattler, U. 2001. Ontology reasoning in the shoq(d) description logic. In *IJCAI*, 199–204.

Huang, Z., and Stuckenschmidt, H. 2005. Reasoning with multi-version ontologies: A temporal logic approach. In *ISWC*, 398–412.

Küsters, R. 2001. *Non-Standard Inferences in Description Logics*, volume 2100 of *LNCS*. Springer.

Lécué, F. 2012. Diagnosing changes in an ontology stream: A dl reasoning approach. In *AAAI*.

Li, L., and Horrocks, I. 2003. A software framework for matchmaking based on semantic web technology. In *WWW*, 331–339.

McGuinness, D. L., and Borgida, A. 1995. Explaining subsumption in description logics. In *IJCAI (1)*, 816–821.

Noia, T. D.; Sciascio, E. D.; Donini, F. M.; and Mongiello, M. 2003. Abductive matchmaking using DLs. In *IJCAI*, 337–342.

Noia, T. D.; Sciascio, E. D.; and Donini, F. M. 2007. Semantic matchmaking as non-monotonic reasoning: A description logic approach. *J. Artif. Intell. Res. (JAIR)* 29:269–307.

Noy, N. F., and Musen, M. A. 2002. Promptdiff: A fixed-point algorithm for comparing ontology versions. In *AAAI/IAAI*, 744–750.

OWL Working Group, W. 2009. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation.

Parsia, B.; Sirin, E.; and Kalyanpur, A. 2005. Debugging owl ontologies. In *WWW*, 633–640.

Sheth, A. P. 2010. Computing for human experience: Semantics-empowered sensors, services, and social computing on the ubiquitous web. *IEEE Internet Computing* 14(1):88–91.

Torta, G., and Torasso, P. 2003. Automatic abstraction in component-based diagnosis driven by system observability. In *IJCAI*, 394–402.