

Computing Best Ontology Excerpts via Weighted Partial Max-SAT Solving*

Jieying Chen¹, Yue Ma¹, Dirk Walther²

¹ Laboratoire de Recherche en Informatique, Université Paris-Sud, France

² Fraunhofer IVI, Dresden, Germany

{jieying.chen,yue.ma}@lri.fr, dirk.walther@ivi.fraunhofer.de

Abstract. We consider the problem of computing best excerpts of ontologies, which are selections of a certain, small number k of axioms that best capture the knowledge regarding a given set of weighted terms. Such excerpts can be useful for ontology selection, for instance. The weights of terms are to reflect different importances of considered terms, for which we propose an incompleteness measure of excerpts based on the term weights and the notion of logical difference. We present a method to compute best k -excerpts by finding all subsumption justifications and solving a weighted partial Max-SAT problem. We demonstrate the viability of our approach with an evaluation on biomedical ontologies.

1 Introduction

In the presence of an ever growing amount of information, organizations and human users need to be able to focus on certain key pieces of information and to intentionally ignore all other possibly relevant parts. Knowledge about complex systems that is represented in ontologies yields collections of axioms that are too large for human users to browse, let alone to comprehend or reason about it. To address the need for ontology summary, the notion of an ontology excerpt was introduced [3]. An ontology excerpt is a fixed-size subset of an ontology, consisting of the most relevant axioms for a given set of terms. These axioms preserve as much as possible the knowledge about the considered terms described in the ontology.

To evaluate the quality of ontology excerpts, we define a semantics-based measure, using Logical Difference [6], to quantify how much semantic meaning is preserved in an excerpt compared to the original ontology. We take the logical difference to be the set of queries that are relevant to an application domain and that produce different answers when evaluated over the considered ontologies. We consider concept subsumption queries over the terms of interest only.

Using an exhaustive search to find the excerpts of an ontology that best preserve the semantic information w.r.t. the ontology is futile as it involves computing all (i.e. exponentially many) subsets of the ontology. In [3] efficient excerpt extraction techniques are investigated that are based on methods stemming from

* This work is partially funded by the ANR project GoAsQ (ANR-15-CE23-0022).

the area of information retrieval [10], a research area which is generally concerned with developing techniques to extract the “most relevant” documents for a query from large data sources. However, these techniques do not yield the best excerpts possible. In this paper, we present an algorithm that computes a best excerpt of an ontology, i.e., there is no subset of the ontology of the same size that better preserves the knowledge about the selected terms. This paper extends the approach from [4] using an encoding to a weighted partial Max-SAT problem to handle weighted signatures.

We proceed as follows. We start in Section 2 with reviewing the notions of logical difference together with difference witness and best excerpt as well as the weighted partial Max-SAT problem. A best k -excerpt of a TBox is a sub-ontology that containing at most k axioms that best capture the knowledge about user’s interests in terms of weighted signature. Moreover, we introduce the notion of subsumption modules of a TBox for a given set of terms. Taking the union of some subsumption modules can yield best excerpts for a weighted signature, as shown in Section 3. We encode the problem of selecting the right subsumption modules on weighted signature as a weighted partial Max-SAT problem. Overall we obtain an algorithm for computing best k -excerpts on weighted signature. In Section 4, we propose a method to rank axioms in excerpts when presenting user a best excerpt. We evaluate our algorithm by using it with the prominent bio-medical ontologies Snomed CT in Section 5. We demonstrate computing best k -excerpts is a viable approach for the task of summarizing large ontologies with a controllable small number of axioms that are most relevant for a given set of terms. We conclude the paper in Section 6.

2 Preliminaries

We briefly recall basic notions related to the ontology excerpts, and the logical difference between terminologies [6, 8]. We assume familiarity with Description Logic; see [1] for a detailed introduction.

2.1 Ontology Excerpt

In this section, we review the notions of a k -excerpt, an incompleteness measure and a best k -excerpt of a TBox [3]. We recall that a signature is a finite set of concept and role names, and we use $\text{sig}(\xi)$ to denote the signature of an syntactic object ξ .

A k -excerpt of an ontology is a subset of the ontology containing no more than k axioms. An incompleteness measure quantifies how much of the knowledge of a TBox about the terms in a signature is captured (or how much is missing) by an excerpt.

Definition 1 (Incompleteness Measure). *Let \mathcal{T} be a TBox, and let Σ be a signature. An incompleteness measure $\mu_{\mathcal{T}}^{\Sigma}$ for \mathcal{T} and Σ is a function that maps every subset $\mathcal{E} \subseteq \mathcal{T}$ to a non-negative real number $\mu(\mathcal{E})$.*

A best k -excerpt best captures the knowledge about Σ in \mathcal{T} among all k -excerpts. How well a k -excerpt captures knowledge is measured by a suitable incompleteness measure. Note that best k -excerpts are generally not unique.

Definition 2 (Best Excerpt). *Let \mathcal{T} be a TBox, let Σ be a signature, and let $k > 0$ be a natural number. Additionally, let $\mu_{\mathcal{T}}^{\Sigma}$ be an incompleteness measure for \mathcal{T} and Σ . A best k -excerpt of \mathcal{T} w.r.t. Σ under $\mu_{\mathcal{T}}^{\Sigma}$ is a k -excerpt \mathcal{E} of \mathcal{T} such that*

$$\mu_{\mathcal{T}}^{\Sigma}(\mathcal{E}) = \min\{\mu_{\mathcal{T}}^{\Sigma}(\mathcal{E}') \mid \mathcal{E}' \text{ is a } k\text{-excerpt of } \mathcal{T}\}.$$

In [3,4], the incompleteness measure μ was set to be the number $|\text{cWtn}_{\Sigma}(\mathcal{T}, \mathcal{E})|$ of concept inclusion difference witnesses in Σ w.r.t. \mathcal{T} , which will be introduced in the next section.

2.2 Logical Difference

We now recall basic notions related to the logical difference between two $\mathcal{EL}\mathcal{H}^r$ -terminologies for \mathcal{EL} -inclusions over a given signature as query language [6, 8].

Definition 3 (Logical Difference). *Let \mathcal{T}_1 and \mathcal{T}_2 be two \mathcal{EL} -terminologies, and let Σ be a signature. The concept inclusion difference between \mathcal{T}_1 and \mathcal{T}_2 w.r.t. Σ is the set $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$ of all \mathcal{EL} -inclusions α of the form $C \sqsubseteq D$ for \mathcal{EL} -concepts C and D such that $\text{sig}(\alpha) \subseteq \Sigma$, $\mathcal{T}_1 \models \alpha$, and $\mathcal{T}_2 \not\models \alpha$.*

In case two terminologies \mathcal{T}_1 and \mathcal{T}_2 are logically different, the set $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$ consists of infinitely many concept inclusions. The *primitive witnesses theorems* from [6] allow us to consider only certain inclusions of a simpler syntactic form.

Theorem 1. *Let \mathcal{T}_1 and \mathcal{T}_2 be two \mathcal{EL} -terminologies, and let Σ be a signature. If $\alpha \in \text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$, then either $A \sqsubseteq D$ or $C \sqsubseteq A$ is a member of $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$, where $A \in \text{sig}(\alpha)$ is a concept name, and C, D are \mathcal{EL} -concepts occurring in α .*

We use the primitive witnesses theorem to obtain a succinct representation of the set $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$.

Definition 4 (Primitive Witnesses). *Let \mathcal{T}_1 and \mathcal{T}_2 be two \mathcal{EL} -terminologies, and let Σ be a signature. We say that concept inclusion difference witnesses in Σ w.r.t. \mathcal{T}_1 and \mathcal{T}_2 are concept names contained in Σ that occur on the left-hand side of inclusions of the form $A \sqsubseteq C$ in $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$ or on the right-hand side of inclusions of the form $D \sqsubseteq A$ in $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$. The set of all such witnesses will be denoted by $\text{cWtn}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$.*

The set $\text{cWtn}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$ is finite as Σ is finite. In particular, it holds that $\text{cDiff}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2) = \emptyset$ iff $\text{cWtn}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2) = \emptyset$.

For (possibly cyclic) \mathcal{EL} -terminologies \mathcal{T}_1 and \mathcal{T}_2 , the set $\text{cWtn}_{\Sigma}(\mathcal{T}_1, \mathcal{T}_2)$ can be computed using the current version of the CEX tool [7] implementing a proof-theoretic approach to the logical difference problem using hypergraphs as introduced in [5] and further extended in [9].

2.3 Weighted Partial Max-SAT problem

Let Π be a countably infinite set of propositional variables which we denote with p, q , etc. The set Φ of propositional logic formulas φ is built according to the following grammar rule: $\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \varphi \leftrightarrow \varphi$ with $p \in \Pi$.

A propositional valuation $v: \Pi \rightarrow \{0, 1\}$ is function mapping a propositional variable $p \in \Pi$ to a truth value $v(p) \in \{0, 1\}$. A propositional formula $\varphi \in \Phi$ is satisfied in v iff $v(\varphi) = 1$. We say that φ is satisfiable iff there exists a valuation v such that $v(\varphi) = 1$.

A literal ℓ is a propositional variable p or a negated propositional variable $\neg p$, and a clause is a disjunction $\ell_1 \vee \dots \vee \ell_n$ of literals ℓ_i ($0 \leq i \leq n$). For a propositional formula $\varphi \in \Phi$, we denote with $\text{Cls}(\varphi)$ the transformation of φ into an equi-satisfiable clause set, i.e. $\text{Cls}(\varphi)$ is a finite set of propositional clauses such that φ is satisfiable iff the formula $\bigwedge_{\psi \in \Psi} \psi$ is satisfiable, where $\Psi = \text{Cls}(\varphi)$. It is known that every propositional formula $\varphi \in \Phi$ can be transformed in polynomial time (in the size of φ) into an equi-satisfiable set of clauses. Note that the clauses in $\text{Cls}(\varphi)$ may contain propositional variables that are not present in φ .

The Boolean Satisfiability Problem (SAT), also known as Propositional Satisfiability Problem, is the problem of deciding whether there exists a valuation that satisfies a given propositional formula φ , or alternatively, all the clauses contained in a set of clauses $\Psi = \text{Cls}(\varphi)$. As an extension, the set Ψ is split into two sets H and S , called *hard* and *soft* clause sets. The problem of finding a valuation that satisfies all the clauses in H and a maximal number clauses in S is called the partial Max-SAT problem.

A weighted clause is a pair (ψ, w_ψ) , where ψ is a clause and w_ψ is a natural number or infinity meaning the cost for falsifying the clause ψ . If a clause $\psi \in H$, then w_ψ is infinite. Given an assignment I and a multiset of weighted clauses Φ , the cost of assignment I on Φ is the sum of the weights of the clauses falsified by I . The Weighted Partial Max-SAT problem for weighted clauses is the problem of finding an optimal assignment to the variables of Φ that minimizes the cost of the assignment on Φ .

2.4 Subsumption Justifications

Subsumption module and justification are defined to capture the set of axioms required to preserve the logical consequences related to a particular concept name [4]. An algorithm for computing subsumption justifications for \mathcal{ELH} -terminologies is presented in [4].

Definition 5 (Subsumption Module & Justification). *Let \mathcal{T} be a terminology, let Σ be a signature, and let $X \in \mathbf{N}_C$. Additionally, let $\mathcal{M} \subseteq \mathcal{T}$.*

We say that \mathcal{M} is an $\langle X, \Sigma \rangle$ -subsumption module of \mathcal{T} iff for every $C, D \in \mathcal{EL}_\Sigma$: $\mathcal{T} \models X \sqsubseteq C$ implies $\mathcal{M} \models X \sqsubseteq C$. Plus $\mathcal{T} \models D \sqsubseteq X$ implies $\mathcal{M} \models D \sqsubseteq X$.

An $\langle X, \Sigma \rangle$ -subsumption justification \mathcal{M} of \mathcal{T} is an $\langle X, \Sigma \rangle$ -subsumption module of \mathcal{T} that is minimal w.r.t. \subseteq .

Property 1. Let \mathcal{T} be a terminology, and let Σ be a signature. Additionally, let \mathcal{M} be a $\langle X, \Sigma \rangle$ -subsumption justification of \mathcal{T} . Then $X \notin \text{cDiff}_\Sigma(\mathcal{T}, \mathcal{M})$.

3 Computing Best k -Excerpts for Weighted Signatures

Best k -excerpts provide users with a way to extract the most relevant axioms from a big ontology with respect to a given signature. However, in certain cases, concepts or roles in the signature can have different importances. To this end, we propose a new incompleteness measure considering a weighted signature and study the corresponding best k -excerpts.

First, we introduce the definition of weighted signature.

Definition 6 (Weighted Signature). A weight function is a function $w : \mathbb{N}_C \cup \mathbb{N}_R \rightarrow \mathbb{R}$ mapping a symbol $\sigma \in \mathbb{N}_C \cup \mathbb{N}_R$ to a value $w(\sigma) \in \mathbb{R} \cup \{\infty\}$. Given a signature Σ and a weight function w , a weighted signature Σ_w is a set $\Sigma_w = \{(\sigma, w(\sigma)) \mid \sigma \in \Sigma\}$.

A signature Σ is *weighted uniformly* under a weight function w iff w assigns the same value to every symbol in Σ , i.e. $w(\sigma_1) = w(\sigma_2)$ for every $\sigma_1, \sigma_2 \in \Sigma$. For a weight function w whose range equals $\{1.0\}$ (i.e., w maps every symbol to 1.0), we also write Σ° instead of Σ_w . For instance, $\Sigma_w = \{(A, 2.0), (B, 0.9), (r, 1.3)\}$ is the weighted signature for the signature $\Sigma = \{A, B, r\}$ and a weight function w satisfying $w(A) = 2.0$, $w(B) = 0.9$ and $w(r) = 1.3$. On the other hand, a weight function w' satisfying $w'(A) = w'(B) = w'(r) = 1.4$ weighs the symbols in Σ uniformly resulting in $\Sigma_{w'} = \{(A, 1.4), (B, 1.4), (r, 1.4)\}$, whereas $\Sigma^\circ = \{(A, 1.0), (B, 1.0), (r, 1.0)\}$.

For best k -excerpts under a weighted signature Σ_w , we use, as an incompleteness measure $\mu_{\mathcal{T}}^{\Sigma_w}$ for Σ_w and \mathcal{T} , the sum of the weights of the \mathcal{EL} -concept inclusion difference witnesses in Σ w.r.t. \mathcal{T} and \mathcal{E} , formally:

$$\mu_{\mathcal{T}}^{\Sigma_w}(\mathcal{E}) = \sum_{X \in \text{cWtn}_\Sigma(\mathcal{T}, \mathcal{E})} w(X)$$

For computing the best k -excerpts for a weighted signature, we present an encoding of the problem to a weighted partial Max-SAT problem, with the aim of delegating the task of finding the best excerpt to a Max-SAT solver. In that way we can leverage the decades of research efforts dedicated to developing efficient SAT solvers for our problem setting.

Note that our encoding requires that a cover of the $\langle A, \Sigma \rangle$ -subsumption justifications has already been computed for every concept name $A \in \Sigma$. Moreover our partial Max-SAT encoding is language agnostic, i.e., it works for any TBox as long as its subsumption justifications can be computed.

As we are interested in finding a subset $\mathcal{M} \subseteq \mathcal{T}$ of a TBox \mathcal{T} that fulfills certain properties, we encode the presence of an axiom $\alpha \in \mathcal{T}$ in \mathcal{M} using a propositional variable p_α . Intuitively, in a solution v to our partial Max-SAT problem we will have that $v(p_\alpha) = 1$ iff the axiom α is contained in the best k -excerpt.

Next we introduce the formulas $F_{\mathcal{M}}$ and $G_{\mathbb{M}}$ with $\mathcal{M} \subseteq \mathcal{T}$ and $\mathbb{M} \subseteq 2^{\mathcal{T}}$ that we will use to encode the covers \mathbb{M} of $\langle A, \Sigma \rangle$ -subsumption justifications \mathcal{M} .

Definition 7 (Encoding of TBox Subsets). *Let \mathcal{T} be a terminology. For every axiom $\alpha \in \mathcal{T}$, let $p_{\alpha} \in \Pi$ be a fresh propositional variable.*

For a subset $\mathcal{M} \subseteq \mathcal{T}$, we define the propositional formula $F_{\mathcal{M}}$ that encodes \mathcal{M} as $F_{\mathcal{M}} := \bigwedge_{\alpha \in \mathcal{M}} p_{\alpha}$. Moreover, for a set $\mathbb{M} \subseteq 2^{\mathcal{T}}$, we define the propositional formula $G_{\mathbb{M}}$ associated with \mathbb{M} as

$$G_{\mathbb{M}} := \bigvee_{\mathcal{M} \in \mathbb{M}} F_{\mathcal{M}} = \bigvee_{\mathcal{M} \in \mathbb{M}} \left(\bigwedge_{\alpha \in \mathcal{M}} p_{\alpha} \right).$$

Example 1. Let $\mathcal{T} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$, $\mathcal{M} = \{\alpha_1, \alpha_2, \alpha_3\}$, and $\mathbb{M} = \{\{\alpha_2, \alpha_3\}, \{\alpha_1, \alpha_4\}\}$. Then $F_{\mathcal{M}} = p_{\alpha_1} \wedge p_{\alpha_2} \wedge p_{\alpha_3}$ and $G_{\mathbb{M}} = (p_{\alpha_2} \wedge p_{\alpha_3}) \vee (p_{\alpha_1} \wedge p_{\alpha_4})$.

We can now define our encoding of the best k -excerpt problem for weighted signature into weighted partial Max-SAT.

Definition 8 (Encoding of the Best Excerpt Problem for Weighted Signature). *Let \mathcal{T} be an \mathcal{EL} -TBox, let Σ_w be a weighted signature, and let $0 \leq k \leq |\mathcal{T}|$. Additionally, for every $A \in \Sigma_w$, let \mathbb{M}_A be a cover of the $\langle A, \Sigma_w \rangle$ -subsumption justifications of \mathcal{T} , and let q_A be a fresh propositional variable.*

The weighted partial Max-SAT problem for finding best k -excerpts of \mathcal{T} w.r.t. Σ_w , denoted with $P_k(\mathcal{T}, \Sigma_w)$, is defined as follows.

We set $P_k(\mathcal{T}, \Sigma_w) := (H_k(\mathcal{T}), S_k(\mathcal{T}, \Sigma_w))$, where

$$H_k(\mathcal{T}) := \text{Card}(\mathcal{T}, k) \cup \bigcup_{A \in \Sigma_w \cap \mathbf{N}_{\mathcal{C}}} \text{Cls}(q_A \leftrightarrow G_{\mathbb{M}_A}),$$

$$S_k(\mathcal{T}, \Sigma_w) := \{ (q_A, w(A)) \mid A \in \Sigma_w \cap \mathbf{N}_{\mathcal{C}} \},$$

and $\text{Card}(\mathcal{T}, k)$ is a set of clauses specifying that at most k clauses from the set $\{p_{\alpha} \mid \alpha \in \mathcal{T}\}$ must be satisfied.

In the hard part of our weighted partial Max-SAT problem, the clauses in $\text{Card}(\mathcal{T}, k)$ specify that the cardinality of the resulting excerpt $\mathcal{E} \subseteq \mathcal{T}$ must be equal to k . We do not fix a certain encoding that should be used to obtain $\text{Card}(\mathcal{T}, k)$, but we note that there exist several techniques that require a polynomial number of clauses in k and in the size of \mathcal{T}^1 (see e.g. [11]). Moreover, for every concept name $A \in \Sigma_w$, the variable q_A is set to be equivalent to the formula $G_{\mathbb{M}_A}$, i.e. q_A will be satisfied in a valuation iff the resulting excerpt will not have A as a difference witness ($A \notin \text{cWtn}_{\Sigma}(\mathcal{T}, \mathcal{E})$).

In the soft part of the problem, the set $S_k(\mathcal{T}, \Sigma_w)$ specifies that the sum of the weights of the concept names in the signature must be maximal by a solution of the problem, enforcing that the resulting excerpt \mathcal{E} will yield the smallest

¹ In our implementation we used the encodings for cardinality constraints that are provided by the Max-SAT solver Sat4j.

possible sum of weights of difference witnesses (whilst obeying the constraint that $|\mathcal{E}| = k$).

We can now show the correctness of our encoding, i.e. a best k -excerpt for a weighted signature can be exactly obtained from any solution to the weighted partial Max-SAT problem $P_k(\mathcal{T}, \Sigma_w)$.

Theorem 2. *Let \mathcal{T} be an \mathcal{EL} -TBox, and let Σ_w be a weighted signature. Then v is a solution of the weighted partial Max-Sat problem $P_k(\mathcal{T}, \Sigma_w)$ if and only if $\{\alpha \in \mathcal{T} \mid v(p_\alpha) = 1\}$ is a best k -excerpt of \mathcal{T} w.r.t. Σ_w .*

Our algorithm for finding a best k -excerpt given a TBox \mathcal{T} and a weighted signature Σ_w (and k with $0 \leq k \leq |\mathcal{T}|$) as input is shown in Algorithm 1. First, the algorithm checks whether $k = 0$ or $k = |\mathcal{T}|$, in which case the empty set or \mathcal{T} itself is the best k -excerpt, respectively. Otherwise, the algorithm continues by computing subsumption justifications for each concept name A in Σ_w and then transfer them to propositional formula $G_{\mathcal{J}_A}$. After the iteration over all the concept names A in Σ_w is complete, the partial Max-SAT problem $P_k(\mathcal{T}, \Sigma_w)$ is constructed with the help of the formulas $G_{\mathbb{M}_A}$ that are stored in S . Subsequently, a solution v of $P_k(\mathcal{T}, \Sigma_w)$ is computed using a partial Max-SAT solver and the best k -excerpt is returned by analyzing which variables p_α have been set to 1 in the valuation v .

Algorithm 1: Computing Best k -Excerpts of \mathcal{T} for Weighted Signature Σ

```

1 function ComputeBestExcerpt( $\mathcal{T}, \Sigma_w, k$ )
2   if  $k = 0$  then
3     | return  $\emptyset$ 
4   if  $k = |\mathcal{T}|$  then
5     | return  $\mathcal{T}$ 
6    $S := \emptyset$ 
7   for every  $A \in \Sigma_w \cap \mathsf{Nc}$  do
8     | Compute  $(A, \Sigma_w)$ -subsumption justifications of  $\mathcal{T}$ :  $\mathcal{J}_\mathcal{T}(A, \Sigma_w)$ 
9     | Transfer  $\mathcal{J}_\mathcal{T}(A, \Sigma_w)$  to its propositional formula  $G_{\mathcal{J}_A}$ 
10    |  $S := S \cup \{G_{\mathcal{J}_A}\}$ 
11  Compute  $P_k(\mathcal{T}, \Sigma_w)$  using  $S$ 
12  Find the set of solutions  $V$  of  $P_k(\mathcal{T}, \Sigma_w)$  using weighted partial Max-SAT solver
13  return  $\{\alpha \in \mathcal{T} \mid v(p_\alpha) = 1, v \in V\}$ 

```

Our algorithm first computes subsumption justifications, which requires exponential time in the size of \mathcal{T} and Σ_w . Then the weighted partial Max-SAT encoding runs in polynomial time. Deciding whether there is a solution of an instance of a Max-SAT problem is NP-complete. Overall we have that Algorithm 1 runs in exponential time in the size of \mathcal{T} and Σ_w in the worst case.

Example 2. Let $\mathcal{T} = \{\alpha_1 := A \sqsubseteq X_1 \sqcap X_2 \sqcap X_3 \sqcap X_4, \alpha_2 := X_1 \sqsubseteq B_1, \alpha_3 := X_2 \sqsubseteq B_1, \alpha_4 := X_3 \sqsubseteq B_2, \alpha_5 := X_4 \sqsubseteq B_2\}$ be a TBox. Let $\Sigma = \{A, B_1, B_2\}$ be a signature. Then, for $X \in \Sigma$, we obtain the following $\langle X, \Sigma \rangle$ -subsumption justifications \mathbb{M}_X : $\mathbb{M}_A = \{\{\alpha_1, \alpha_2\}, \{\alpha_1, \alpha_3\}, \{\alpha_1, \alpha_4\}, \{\alpha_1, \alpha_5\}\}$, $\mathbb{M}_{B_1} = \{\{\alpha_1, \alpha_2\}, \{\alpha_1, \alpha_3\}\}$ and $\mathbb{M}_{B_2} = \{\{\alpha_1, \alpha_4\}, \{\alpha_1, \alpha_5\}\}$. Each subsumption justification \mathbb{M}_X is encoded as a propositional logic formula \mathcal{G}_X as follows:

$\mathcal{G}_A = ((p_1 \wedge p_2) \vee (p_1 \wedge p_3) \vee (p_1 \wedge p_4) \vee (p_1 \wedge p_5))$, $\mathcal{G}_{B_1} = ((p_1 \wedge p_2) \vee (p_1 \wedge p_3))$ and $\mathcal{G}_{B_2} = ((p_1 \wedge p_4) \vee (p_1 \wedge p_5))$.

Now let w_1, w_2, w_3 be weight functions such that $w_1(A) = w_1(B_1) = w_1(B_2) = 1.0$, $w_2(A) = w_2(B_1) = 1.0$ and $w_2(B_2) = 2.0$, and $w_3(A) = w_3(B_2) = 1.0$, $w_3(B_1) = 2.0$. Feeding the resulting weighted signatures Σ_{w_1} , Σ_{w_2} and Σ_{w_3} together with $k = 2$ to a SAT solver yields the following solutions. For Σ_{w_1} , we obtain four assignments of truth values for the propositional variables in the tuple $\langle p_1, p_2, p_3, p_4, p_5 \rangle$ (in this order): $\langle 1, 1, 0, 0, 0 \rangle$, $\langle 1, 0, 1, 0, 0 \rangle$, $\langle 1, 0, 0, 1, 0 \rangle$ and $\langle 1, 0, 0, 0, 1 \rangle$. Then we decode the assignments to sets of axioms. We obtain the following best 2-excerpts for Σ_{w_1} : $\{\alpha_1, \alpha_2\}$, $\{\alpha_1, \alpha_3\}$, $\{\alpha_1, \alpha_4\}$ and $\{\alpha_1, \alpha_5\}$.

In the case of Σ_{w_2} , however, the solutions returned by the SAT-solver are the assignments $\langle 1, 0, 0, 1, 0 \rangle$ and $\langle 1, 0, 0, 0, 1 \rangle$. Consequently, the best 2-excerpts for Σ_{w_2} are $\{\alpha_1, \alpha_4\}$ and $\{\alpha_1, \alpha_5\}$.

Finally, for Σ_{w_3} , the SAT-solver yields $\langle 1, 1, 0, 0, 0 \rangle$ and $\langle 1, 0, 0, 1, 0 \rangle$ resulting in $\{\alpha_1, \alpha_2\}$ and $\{\alpha_1, \alpha_4\}$ as the best 2-excerpts for Σ_{w_3} .

Example 2 shows that, different weights of Σ -symbols can influent the final results of best excerpts. By giving higher weights on preferring Σ -concept names (like Σ_{w_2} or Σ_{w_3}), users can reduce the number of best k-excerpts. However, in some extreme cases, as shown in the following example where a signature has a much higher weight then the others, the best excerpts for the wighted signature can be totally irrelevant to those w.r.t. Σ° .

Example 3. Let $\mathcal{T} = \{\alpha_1 := A_1 \sqsubseteq X, \alpha_2 := X \sqsubseteq Y, \alpha_3 := Y \sqsubseteq Z, \alpha_4 := A_2 \sqsubseteq B_1, \alpha_5 := A_3 \sqsubseteq B_2\}$ be a TBox. Let $\Sigma = \{A_1, A_2, A_3, B_1, B_2, Y, Z\}$ be a signature. Now let w_1, w_2 be weight functions such that $w_1(A_1) = w_1(A_2) = w_1(A_3) = w_1(B_1) = w_1(B_2) = w_1(Y) = w_1(Z) = 1.0$ (i.e., Σ_{w_1} is Σ°) and $w_2(A_1) = w_2(A_2) = w_2(A_3) = w_2(Y) = 1.0$ and $w_2(Z) = 9.0$. We obtain that the best 3-excerpt $\mathcal{E}_{\Sigma^\circ}$ for Σ° is $\{\alpha_4, \alpha_5\}$ ². The best 3-excerpt $\mathcal{E}_{\Sigma_{w_2}}$ for Σ_{w_2} is $\{\alpha_1, \alpha_2\}$. However, $|\text{cWtn}_\Sigma(\mathcal{T}, \mathcal{E}_{\Sigma^\circ})| = |\{A_1, X, Y\}| = 3$ and $|\text{cWtn}_\Sigma(\mathcal{T}, \mathcal{E}_{\Sigma_{w_2}})| = |\{A_2, A_3, B_1, B_2\}| = 4$.

In Example 3, as the $w_2(Z)$ is much higher than the rest concept names in signature, $\mathcal{E}_{\Sigma^\circ}$ is not the same as \mathcal{E}_{Σ_w} . The following theorem provides a guide of how to restrict weight function in order to keep monotonicity of best excerpts even though w changes.

Proposition 1. *Let \mathcal{T} be an \mathcal{EL} -TBox, Σ_w be a weighted signature and let Σ° be a uniformed signature under the weighted function w .*

- (i) *If $1 \leq w(\sigma) \leq 1 + \frac{1}{|\Sigma \cap \mathbf{N}_C|}$ for every $\sigma \in \Sigma \cap \mathbf{N}_C$, then every best k-excerpt of \mathcal{T} w.r.t. Σ_w will be a best k-excerpt of \mathcal{T} w.r.t. Σ° , but not vice versa.*
- (ii) *If $\sum_{\sigma_i \in \Sigma_1} w_{\sigma_i} > \sum_{\sigma_j \in \Sigma_2} w_{\sigma_j}$ for any $\Sigma_1, \Sigma_2 \in 2^{\Sigma \cap \mathbf{N}_C}$ and $|\Sigma_1| > |\Sigma_2|$, iff every best k-excerpt of \mathcal{T} w.r.t. Σ_w is a best k-excerpt of \mathcal{T} w.r.t. Σ° .*

² The best 3-excerpt $\mathcal{E}_{\Sigma^\circ}$ for Σ_{w_1} only contains two axioms α_4 and α_5 , as for any best 3-excerpt \mathcal{E}' , $\mu_{\mathcal{T}}^{\Sigma^\circ}(\mathcal{E}') = \mu_{\mathcal{T}}^{\Sigma^\circ}(\{\alpha_4, \alpha_5\})$

4 Ranking Axioms in Best Excerpts

Best excerpts are useful to work with when a whole ontology exceeds the capacity of a system or a human user. However, multiple axioms often exist in the best excerpts. To better assist users, it would be desirable to give a ranking of these axioms according to their importances. In this section we present an approach to weight axioms according to weighted signatures.

To achieve a ranking, we introduce the definition of weighted axioms.

Definition 9 (Weighted Axiom). *An axiom weight function ϕ is a function $\phi : \mathcal{T} \rightarrow \mathbb{R}^+$ mapping an axiom $\alpha \in \mathcal{T}$ to a value $\phi(\alpha) \in \mathbb{R}^+$.*

Given an axiom weight function, it is then natural to introduce an axiom ranking. Intuitively, an axiom with a larger weight should be ranked more highly.

Definition 10 (Axiom Ranking). *Given an axiom weight function ϕ and two axioms α and β , we say α is preferred than β , denoted $\alpha \succeq \beta$, if $\phi(\alpha) \geq \phi(\beta)$.*

To achieve an axiom ranking, in the following, we show that a signature weight function can induce an axiom weight function. Intuitively, an axiom that appears in subsumption justifications of more Σ -symbols with higher weights should be more highly weighted.

Definition 11. *Let \mathcal{T} be a TBox, Σ_w a signature with the weight function w . Additionally, \mathbb{M}_σ be the set of all subsumption justifications of σ . Then the axiom weight function $\phi(\cdot)$ induced by w is:*

$$\phi_{\Sigma_w}(\alpha) = \sum_{\{\sigma \in \Sigma_w \mid \exists M \in \mathbb{M}_\sigma \text{ s.t. } \alpha \in M\}} w(\sigma)$$

We denote \succeq_{Σ_w} the axiom ranking induced by ϕ_{Σ_w} .

Example 4 (Example 2 contd.). Note that α_1 (resp. $\alpha_2, \alpha_3, \alpha_4, \alpha_5$) is contained in at least one of the subsumption justifications of Σ -symbols $\{A, B_1, B_2\}$ (resp. $\{A, B_1\}, \{A, B_1\}, \{A, B_2\}, \{A, B_2\}$). So for w_3 , $\phi(\alpha_1) = w(A) + w(B_1) + w(B_2) = 4.0$, $\phi(\alpha_2) = \phi(\alpha_3) = w(A) + w(B_1) = 2.0$, and $\phi(\alpha_4) = \phi(\alpha_5) = w(A) + w(B_2) = 3.0$. Hence, for the 2-best excerpts w.r.t. w_3 : $\{\alpha_1, \alpha_2\}$ and $\{\alpha_1, \alpha_4\}$, we can order the axioms by $\alpha_1 \succeq_{\Sigma_w} \alpha_2$ and $\alpha_1 \succeq_{\Sigma_w} \alpha_4$.

It is easy to see that the axiom function $\phi(\cdot)$ in Definition 11 gives a total ordering over the axioms in \mathcal{T} w.r.t. to a given weighted signature Σ_w . Moreover, it is independent of excerpts as stated in the following.

Proposition 2. *Let \mathcal{T} be a TBox, Σ_w a signature with the weight function w , for any best excerpts \mathcal{E}_1 and \mathcal{E}_2 and axioms $\alpha_1, \alpha_2 \in \mathcal{E}_1 \cap \mathcal{E}_2$, $\alpha_1 \succeq \alpha_2$ in \mathcal{E}_1 iff $\alpha_1 \succeq \alpha_2$ in \mathcal{E}_2 .*

That is, \succeq is a stable ordering of the axioms given a weighted signature. However, note that the axioms in best k -excerpts are not monotonic with respect to the size of k (see a counter-example given in [3]).

Time (s)	Snomed CT				NCI			
	Minimal	Maximal	Median	STDEV	Minimal	Maximal	Median	STDEV
MaxSat	0.01	1.27	0.15	0.35	0.01	0.16	0.01	0.02

Table 1. Times for computing best excerpts encoding the input for and running SAT4J

5 Evaluation

To demonstrate that our algorithm works in practise, we implemented a JAVA prototype using SAT4J as a Max-SAT solver [2]. All the tests were performed on a server running Ubuntu 15.10 with Intel Xeon 2.50GHz Core 4 Duo CPU with 64G RAM. We used the prominent medical ontologies Snomed CT (version Jan 2016)³ consisting of 317 891 axioms, and the NCI Thesaurus (NCIt, version 16.03d)⁴ containing 165 341 axioms.

The first experiment is to test the efficiency of our encoding method. Table 1 shows the time needed to compute best excerpts of NCI and Snomed CT for randomly generated signatures of different sizes (10/30/50 concept names and 10/30/50 roles names). In total over 3 000 signatures have been considered. After precomputing the subsumption justifications for every concept name in every signature, the time was measured as that needed to prepare the input and run the partial Max-SAT solver to compute best k -excerpts for every size k . The minimal, maximal, median, and standard deviation of the execution times show that computing best excerpts of any size using Max-SAT is very efficient given that the subsumption justifications have been computed.

In the second experiment, we compare the size of locality based modules with the number of axioms in best excerpts needed to preserve a certain amount of knowledge. We denote with $\#(\text{Preserved}_{\Sigma}(\text{best}) = n)$, for $n \in \{1, 2\}$, the minimal number of axioms needed by a best excerpt to preserve the knowledge of n concept names w.r.t. the signature Σ (i.e., the number k of axioms of a best excerpt \mathcal{E} of \mathcal{T} such that $n = |\Sigma \cap \mathbf{N}_{\mathcal{C}} \setminus \text{cWtn}_{\Sigma}(\mathcal{T}, \mathcal{E})|$). Instead of using random signatures, however, we consider a scenario where a user searches for sub-ontologies of Snomed CT related to a particular concept name. We computed 2 500 different signatures each consisting of a concept name related to diseases, the TOP-concept (called ‘SNOMED CT Concept’) and all role names of Snomed CT. All weights of Σ -symbols are equal in this experiment.

In Figure 1, the 2 500 signatures are presented along the x-axis in ascending order w.r.t. the size of their respective $\perp\top^*$ -local modules (black line). The y-axis represents the number of axioms in the module and excerpts for a signature. The red (resp. green) line presents the sizes of best excerpts that preserve the knowledge for one (resp. two) concept name(s), i.e., $\#\text{Preserved}_{\Sigma}(\text{best}) = 1$ (resp. $\#\text{Preserved}_{\Sigma}(\text{best}) = 2$). Best excerpts provide a conciser way for zooming in on an ontology, in particular, when sacrificing completeness is acceptable. For

³ <http://www.ihtsdo.org/snomed-ct>

⁴ http://evs.nci.nih.gov/ftp1/NCI_Thesaurus

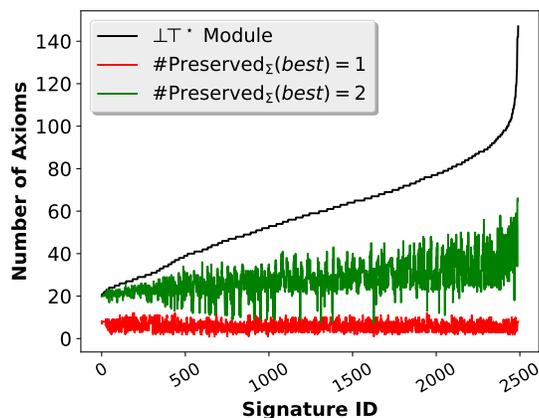


Fig. 1. Comparison of the size of best excerpts preserving one or two concept names with the size of $\perp T^*$ -local modules for 2500 signatures

this experiment, after computing the subsumption justifications of all concept names in a signature, it only took 0.15s on average to compute the best excerpts.

6 Conclusion

We have presented an algorithm for computing a best k -excerpt of a TBox for given weighted signature. Best k -excerpts are k -element subsets of an ontology that best preserve entailments over the target signature compared to the ontology [3]. The measure of incompleteness of the excerpt w.r.t. the ontology is based on the incompleteness measure taking into account weights of concept names in the signature. We have implemented our algorithm and we have presented an experimental evaluation to demonstrate the viability of computing best excerpts of large biomedical ontologies in practice. We leave investigating alternative incompleteness measures for future work.

References

1. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The description logic handbook: theory, implementation, and applications. Cambridge University Press, 2 edn. (June 2010)
2. Berre, D.L., Parrain, A.: The sat4j library, release 2.2. JSAT 7(2-3), 59–6 (2010)
3. Chen, J., Ludwig, M., Ma, Y., Walther, D.: Towards extracting ontology excerpts. In: Proceedings of KSEM’15. pp. 78–89 (2015)
4. Chen, J., Ludwig, M., Ma, Y., Walther, D.: Zooming in on ontologies: Minimal modules and best excerpts. In: Proceeding of ISWC’17. pp. 173–189 (2017)
5. Ecke, A., Ludwig, M., Walther, D.: The concept difference for \mathcal{EL} -terminologies using hypergraphs. In: Proceedings of DChanges’13 (2013)

6. Konev, B., Ludwig, M., Walther, D., Wolter, F.: The logical difference for the lightweight description logic EL. *Journal of Artificial Intelligence Research* 44, 633–708 (2012)
7. Konev, B., Ludwig, M., Wolter, F.: Logical difference computation with CEX2.5. In: *Proceedings of IJCAR'12*. pp. 371–377 (2012)
8. Konev, B., Walther, D., Wolter, F.: The logical difference problem for description logic terminologies. In: *Proceedings of IJCAR'08*. pp. 259–274 (2008)
9. Ludwig, M., Walther, D.: The logical difference for ELHr-terminologies using hypergraphs. In: *Proceedings of ECAI'14*. pp. 555–560 (2014)
10. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press (2008)
11. Sinz, C.: Towards an optimal cnf encoding of boolean cardinality constraints. In: *Proceedings of CP'05*. pp. 827–831 (2005)