Subontology Extraction Using Hyponym and Hypernym Closure on is-a Directed Acyclic Graphs
Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi

To cite this version:
Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi. Subontology Extraction Using Hyponym and Hypernym Closure on is-a Directed Acyclic Graphs. IEEE Transactions on Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers, 2012, 24 (12), pp.2288-2300.

HAL Id: hal-00797150
https://hal.archives-ouvertes.fr/hal-00797150
Submitted on 5 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Sub-ontology extraction using hyponym and hypernym closure on is-a directed acyclic graphs

Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi

Abstract—Ontologies are successfully used as semantic guides when navigating through the huge and ever increasing quantity of digital documents. Nevertheless, the size of numerous domain ontologies tends to grow beyond the human capacity to grasp information. This growth is problematic for a lot of key applications that require user interactions such as document annotation or ontology modification/evolution. The problem could be partially overcome by providing users with a sub-ontology focused on their current concepts of interest. A sub-ontology restricted to this sole set of concepts is of limited interest since their relationships can generally not be explicit without adding some of their hyponyms and hypernyms. This paper proposes efficient algorithms to identify these additional key concepts based on the closure of two common graph operators: the least common-ancestor and greatest common descendant. The resulting method produces ontology excerpts focused on a set of concepts of interest and is fast enough to be used in interactive environments. As an example, we use the resulting program, called OntoFocus (http://www.ontotoolkit.mines-ales.fr), to restrict, in few seconds, the large Gene Ontology (~30,000 concepts) to a sub-ontology focused on concepts annotating a gene related to breast cancer.

Index Terms—[II. Artificial Intelligence / IV. Knowledge Representation Formalisms and Methods]
[II. Discrete Mathematics / II. Graph Theory]
Keywords—Sub-ontology extraction, Ontology transformation, Directed acyclic graph, Least common ancestor, Greatest common descendant.

1 INTRODUCTION

With rapid technological progress, the scope of digital data is overwhelming, ranging from real-time data flows to digitalized documents through specialized databases. Many dedicated applications make use of them, and they often become the core of company decision-making processes. However, their number constitutes a bottleneck for human beings striving to get an overall grasp of information and an impediment to high performance computing. Efficient filtering tools are thus needed to provide the user with the right information at the right time.

Wondering about the Semantic Web, Tim Berners Lee imagined an environment where software agents and humans might cooperate using the same information [1]. Its architecture relies on machine readable semantic models (ontologies) as backbones to structure collections of information and sets of rules to conduct reasoning. This architecture was proposed for the Web, but also appears to be efficient for Information Systems (IS) dealing with large amounts of data [2]. Specialized ontologies endow many Web independent applications with a semantic guide while browsing through documents, filtering and retrieving relevant information or analyzing data in a given context.

However, ontologies, especially those shared and accepted as standards in a given domain, tend to grow beyond the human capacity to manage them (browse, visualize, edit, analyze and share.) This growth is problematic for a lot of applications that require user interactions such as document annotation or ontology modification/evolution. A natural solution to help users would be to provide them with a sub-ontology focused on their current concepts of interest. Of course, as mentioned by [3], individual concepts or terms may always be found by textual searches within the ontology, but links between concepts are tedious to look for. Moreover, given an initial set of concepts of interest, their relationships can generally not be explicit without adding some hyponyms and hypernyms.

In this paper, given a reference-ontology, a “good” sub-ontology is envisaged as the smallest excerpt containing concepts of interest, additional concepts to specify their interconnections and corresponding relationships. We identified (but were not limited to) three main applications:

- Conception/Extension of ontologies: when designing an ontology, it is often necessary to zoom in on a specific part of it. While modifying an ontology, people may consider a very restricted subset of concepts and analyze them through relationships that interconnect them before adding new concepts or new relations or reorganizing them.
- Ontology based information retrieval/filtering: given a set...
of concepts that annotate a set of documents, having relationships that link them may improve the user’s understanding of the annotations.

• Browsing and display of ontologies and related indexations: human cognitive and perceptive limits underline the crucial need for suitable representation tools. Displaying the overall ontology is generally unconceivable and thus some views of it often have to be built.

Sub-ontology is a pre-requisite for ontology visualization which is essential to favor user interactions with ontology based systems. Indeed, we agree with [4] that “regardless of the ontology, presenting terms in a graphical context makes the relationships of ontology terms clear, provides context for annotations, and makes the examination of large annotation sets feasible.”

Some sub-ontology extraction methods have been proposed in the Knowledge Engineering literature. They are discussed in the state of the art presented in section 2. Most of them highlight the importance of the is-a relationship that may be considered as the backbone of the ontology. This kind of relationship can be represented as a directed acyclic graph (dag). Given an initial set of nodes, this paper proposes a definition of relevant nodes using graph operators. Section 3 reviews useful graph definitions and introduces the notion of least common ancestor closure, denoted \( U_{lca} \) closure, to determine the searched subset of relevant concepts. Section 4 presents optimized algorithms to compute this closure. Having grasped the set of relevant concepts, Section 5 details the process for inferring relations among them.

The major contribution of this work is to propose a sub-ontology extraction method that significantly reduces the number of concepts in a very short time (an ontology having thirty thousand concepts may be restricted to a hundred concepts in a few seconds on a standard desktop computer). It favors different displays corresponding to different views on the reference ontology. It is thus possible to embed it within interactive applications where domain experts may focus on relevant concepts of their specialized domain. Visualization of the corresponding ontology excerpt enables users to better comprehend the underlying semantics of their concepts of interest. This extraction algorithm was developed in a program called OntoFocus that takes the OWL reference ontology and a list of concepts of interest as input. Written in JAVA and using the Jena library, this software may be freely used on-line at the address http://www.ontotoolkit.mines-alès.fr/. An application in the life-science domain is detailed in Section 6. The results are analyzed and discussed and outcomes are presented. Section 7 concludes by recapitulating the approach and discussing its impact.

2 STATE OF THE ART: DIFFERENT SUB-ONTOMETRY EXTRACTION APPROACHES

The need for restricted ontologies, particularly in the Life Science domain, is underlined in [5] where different approaches are compared and their strengths and weaknesses are analyzed with respect to the restriction objectives. For authors, ontology restriction methods often involve a two-step process: identification of the set of concepts of interest and then the restriction phase. However, they argue that none can be applied universally. In what follows, the sub-ontology extraction context is given and some methods are presented, organized according to three main application contexts: 1) conception of an ontology, 2) reuse of existing ontologies, and 3) visualization. We then consider sub-ontology extraction in the light of graph theory and present related works in this domain.

2.1 Pruning and trimming ontologies during their conception

Concerning ontology conception, some studies have been conducted to try to automatically identify conceptual spaces from a corpus analysis using natural language processes [6], even during the querying phase [7]. A statistical approach is usually used to determine potential concepts based on their frequency in the corpus and relationships among them are created de novo. After having automatically collected and organized concepts from the textual corpus, such approaches often rely on a pruning and trimming phase [8, 9]. The trimming step proposed in [8] to reduce the built ontology is closely related to sub-ontology extraction. Yet this step is more a way to clean up de novo relationships and remove redundant, useless or too fine-grained potential concepts than a real sub-ontology extraction. One point must be underlined here. Some concepts, identified as belonging to high level ontology or domain ontology, cannot be removed and are always kept in their “sub-ontologies”. We do not distinguish such levels in our approach and all nodes are considered in the same manner.

2.2 Sub-ontology extraction to favor their recycling

An alternative strategy is to reuse parts of existing ontologies. A Web ontology segmentation was proposed in [10] to extract customized ontologies, specific to given applications, from large ontologies. Unfortunately, by keeping all ascendants and descendants of concepts of interest, the size of the resulting sub-ontology might be very large. This customized ontology may be used alone or to enrich a project-specific ontology. In this latter context, partial ontologies are also called modules [11, 12]. When a project related ontology P is enriched with an external module Q, the safety of this enrichment must be ensured. Indeed, if a common label is used for distinct concepts in the two ontologies, these can lead to some erroneous reasoning. More formally, an enrichment of P, using Q, is said to be safe when it “produces exactly the same logical consequences over the vocabulary of Q as Q alone” [11, 12]. This property is not straightforward to check and may result in complex algorithms and high computation times. It should however be noted that safety is only related to ontology merging and not to sub-ontology extraction.

2.3 Sub-ontology extraction as visualization aid

Displaying large ontologies is challenging. This problem can be tackled by two complementary approaches: using infor-
mation visualization techniques suited for scalability (e.g., treemap, hyperbolic trees, etc.) and/or using a semantic filtering based on sub-ontology extraction. A complete overview of ontology visualization tools is provided in [13]. Most of them only propose functionalities to zoom in on a local part of the ontology hierarchy rather than displaying an explicit and semantic excerpt of it. Solutions to obtain such excerpts are discussed in what follows.

When focusing on a subset of concepts (e.g., those indexing a given document or those over represented in the index of a set of documents) a graphical representation of their hierarchical relationships is very helpful. The most widespread solution is to display concepts that are in this subset plus those that are ancestors of one of them. This straightforward solution is (manually) used in many publications (e.g. [4, 14]) as well as within Web-based tools.

An Australian team proposed a solution for what they call the “user-driven automatic extraction of valid sub-ontology” [15, 16]. The need for sub-ontology extraction is clearly highlighted in [16], where the authors point out the fact that an application focuses only on particular aspects of the whole ontology. For the authors, sub-ontologies are valid independent ontologies, known as materialized ontologies, which may contain thousands of elements, [...], the two solutions differ with respect to various points. In their proposal [3], concepts of the materialized ontology may be a compression (rename) of concepts of interest. They are obtained via optimization schemes, a logical grouping of rules, and actions of the user. Conversely, for us, each concept that appears in the sub-ontology is directly derived from the reference ontology. Moreover, they considered the possibility of extending each sub-ontology independently [17], whereas we propose a rapid solution to allow the creation of sub-ontologies on request. The latter may thus be considered as instantaneous views on the large ontology, which is the sole ontology that can be used for automated reasoning (e.g. during information retrieval) and that can evolve. This avoids redundancy and related maintenance difficulties. Moreover, while their method needs to be parallelized for handling large ontologies, ours can filter such ontologies in few seconds on a standard desktop computer.

Having the set of concepts of interest (or signature), our main objective is to identify hyponyms and hypernyms that must be added to clarify relationships among the remaining concepts and thus to obtain an explicit sub-ontology.
than SCA and ASA strategies. Moreover, since it is based on a well defined lca-closure operator, its output does not depend on the arbitrary choice of a spanning tree.

2.5 A new graph based sub-ontology approach

In the following, sub-ontology extraction is tackled by using an algorithmic approach that is based on is-a relationships and two common operators: least common ancestor (lca) and greatest common descendant (gcd). When the reference-ontology is represented by a Direct Acyclic Graph (dag), the “explicit” sub-ontology will be a subgraph that is closed for the above mentioned operators.

Before detailing our solution, some choices have to be justified. As underlined in the introduction, the is-a relationship is of particular interest. It is often considered as the backbone of the ontology – perhaps because it is commonly used by people to organize things around them. This relationship is central to ontologies as it is the sole one that appears in formal ontology definitions ([21] or [22]). Moreover, previous works have been proposed to use the hierarchical domain structure to compute similarity [23] or in information retrieval systems [24, 25]. Within these contexts, adding other relations just adds noise. Therefore we decided to focus on the dag representing is-a relationships to identify relevant concepts. Other relationships are considered and added during post-processing, as described in section 5.3.

Basic filtering, which involves keeping ancestors of concepts of interest, may preserve too many concepts since some concepts may have numerous ancestors (e.g. Aspirin has 152 ancestors in UMLS2003AA). The choice of lca operator to identify relevant concepts is motivated and validated through end-user evaluations in [26], with the aim of assisting experts who have to choose an ancestor concept to summarize a group of concepts for adjusting indexation granularity. The resulting tool displays the whole ancestor hierarchy but emphasizes lca concepts to assist experts in making choices. Our approach goes a step further and filters the ancestor hierarchy to display only ancestral concepts that are least common ancestors of some other preserved concepts. Moreover, unlike existing solutions focused on ancestors, we also preserve greatest common descendant concepts. These two key improvements considerably reduce the number of displayed concepts while preserving enough of them to render relationships among initial concepts. As a counterpart, a straightforward algorithm based on numerous lca computations results in high computation times.

Some works have been proposed to efficiently find the lca of concepts [27, 28]. However, they often calculate lca only for a pair of concepts (and not for a set of them) and do not search for the possibly numerous lcas but only for one of them. The algorithm introduced in [28] for multilabeled trees (i.e. where a single term can label more than one node) can easily be adapted for dag; but this solution cannot be used for large ontologies due to its time complexity. The next section introduces an efficient algorithm to identify relevant concepts based on lca. The key idea is to exploit the dag structure so as to compute in a single graph, traversal lca of any pair of concepts within the extracted sub-ontology.

In the following section, based on the experience of the graph community, a solution is developed to build sub-ontologies.

3 LEAST COMMON ANCESTOR OPERATOR AND ITS CLOSURE

3.1 Preliminary definitions

Restricting a tree to a subset of nodes is a well defined operation used in various computer science fields. Yet its extension to dags is not straightforward. This section introduces some basic graph notations and provides two equivalent formal definitions of lca closure in dags.

Before going any further, and to avoid misunderstandings, it is important to note that in most of the graph literature dag edges are oriented from parents toward children, whereas in the ontology literature edges of the is-a dag are instead oriented from children toward parents in order to stress the is-a semantic. Though the following results hold for any dag, we clearly worked in the ontology framework and thus adopt the corresponding edge orientation convention (from children towards parents).

The following definitions are provided to make the paper self-contained. For further definitions on graphs see [29].

The is-a relationship graph is a dag $G = (V, E)$. The indegree $d^{-}(v)$ of a vertex $v$ is the number of edges with head $v$ (tail $v$). When $G$ contains a directed $(v, u)$-path, the vertex $u$ is said to be an ancestor of $v$ and the vertex $v$ is a descendant of $u$. For a non-empty subset $W$ of $V$, the subgraph of $G$ whose vertex set is $W$ and whose arc set is the set of arcs of $G$ that have both ends in $W$ is called the subgraph of $G$ induced by $W$ and is denoted $G[W]$.

Using these notations, it is now possible to define the least common ancestor operator that is extensively used in our sub-ontology extraction approach. Given a vertex $v$ of the dag $G = (V, E)$, the set $A_{G}(v)$ denotes the subset of ancestors of $v$ in $G$. The generalization of this definition to a set $S \subseteq V$ of vertices is straightforward, i.e. $A_{G}(S) = \bigcap_{v \in S} A_{G}(v)$. For simplicity, we will omit index $G$ from the notations whenever there is no ambiguity.

Definition 3.1.1. ([30]) The least common ancestors lca$(S)$ of a vertex subset $S \subseteq V$ with respect to a dag $G = (V, E)$ are vertices $u \in A(S)$, such that $d^{-}(u) = 0$ in the graph $H = G[A(S)]$ induced by $A(S)$.

This definition generalizes the widely known definition of least common ancestor (see [30-33]) for a couple of vertices, i.e. $lca((x, y)) = lca(x, y)$. It follows immediately from Definition 3.1.1 that $x = lca(x, y)$ if there is a directed $(y, x)$-path. By extension, we define $lca(x, x) = x$ for all $x$.

Note that, unless the dag $G$ is a tree, the existence of a pair of nodes $x, y \in S$ such that $lca(x, y) = lca(S)$ is not guaranteed. For instance, in the example presented in Fig. 1, $lca(C1, C2, C3) = \{A1, A2\}$ while $lca(C1, C2) = \{B1, A2\}$, $lca(C1, C3) = B2$ and $lca(C2, C3) = \{A1, B3\}$.


3.2 \(U_{\text{lca}}\) operator and two equivalent definitions of its closure in a dag

Closure operators are widely used in mathematics, especially in geometry. The best known example concerns convexity in a Euclidian space where a convex polygon can be obtained by giving a finite set of points and the segment operation \(s(x, y)\) (that could correspond to \(\text{lca}(x, y)\) operation).

In the previous section, we provided an intuitive and natural definition of the least common ancestors of a set of vertices denoted \(\text{lca}(S)\). Yet, when considering a set of vertices \(S\), every least common ancestor of a pair of vertices of \(S\) is a key vertex to gain insight into relationships between vertices of \(S\). We thus now introduce a new operator that makes use of the \(\text{lca}(x, y)\) operator and generalizes it to a well defined closure operator.

**Definition 3.2.1.** Let \(S\) be a subset of vertices of \(G\). The \(U_{\text{lca}}\) operator on \(S\) is defined as:

\[
U_{\text{lca}}(S) = \bigcup_{x,y \in S} \text{lca}(x, y) \tag{1}
\]

It follows from this definition that \(S \subseteq U_{\text{lca}}(S)\) and that the \(U_{\text{lca}}\) operator is monotonic, i.e.:

\[
A \subseteq B \Rightarrow U_{\text{lca}}(A) \subseteq U_{\text{lca}}(B).
\]

Fig. 2 illustrates the definition of the \(U_{\text{lca}}\) operator. This example also highlights the difference between the two least common ancestor operators. Given the set \(S = \{C1, C2, C3\}\), \(U_{\text{lca}}(S) = \{A1, A2, B1, B2, B3, C1, C2, C3\}\), while \(\text{lca}(S) = \{A1, A2\}\).

In some cases, having only \(U_{\text{lca}}(S)\) and/or \(\text{lca}(S)\) is not enough to understand all relationships among concepts of \(S\). This case is depicted in Fig. 3, where concept \(A\) is helpful for understanding \(C1\) and \(C2\) relationships but is neither included in \(U_{\text{lca}}(S)\) nor in \(\text{lca}(S)\). Concept \(A\) is of interest since \(A\) is the \(\text{lca}\) of \(B1\) and \(B2\), which in turn are \(\text{lca}\) of two concepts of \(S\). This leads us to the following definition of the \(U_{\text{lca}}\) closure of \(S\).

Let \(S\) be a subset of vertices of \(G\). Let the increasing set sequence defined by: \(S_0 = S\) and \(S_{k+1} = U_{\text{lca}}(S_k), k = 0, 1, ... \). As \(G\) is finite, there is a number \(c, 0 \leq c \leq |V|\) such that \(\forall k \geq c, S_k = U_{\text{lca}}(S_k)\). This fixpoint (or fixset) is reached because of the monotonicity of the \(U_{\text{lca}}\) operator. In fact, once this relation holds for a given \(k\), it holds for all greater values. So \(c\) and \(S_c\) are well defined.

**Definition 3.2.2.** The number \(c\) is called the closure index and the set \(S_c\) is called the \(U_{\text{lca}}\) closure of \(S\) and is denoted \(\hat{S}\).

This definition provides a simple iterative algorithm to compute \(U_{\text{lca}}\) closure. The time complexity of this algorithm is related to the closure index. In the simple case where \(G\) is a tree, the closure index cannot be greater than 1. The following lemma shows that this is no longer true in the general case.

**Lemma 3.2.1.** For a dag \(G = (V, E)\) and a set \(S \subseteq V\), the number of iterations needed to obtain \(\hat{S}\) is \(O(|V|)\).
These definitions provide the framework for our sub-ontology extraction.

4 An Efficient Algorithm to Compute $U_{\text{lca}}$ Closure

Since ontologies may contain several thousand concepts, efficient algorithms are needed to compute the $U_{\text{lca}}$ closure of a set $S$ of concepts of interest with respect to their underlying is-a dag $G = (V, E)$.

4.1 Straightforward algorithms to compute $U_{\text{lca}}$ closure

We give below, as a lower bound on complexity, two straightforward solutions to calculate $U_{\text{lca}}$ closure (the second being a slight improvement of the first, see Algorithm 2).

We denote $s_k = |S_k|$ and estimate the number $N_{\text{lca}}$ of calls of the $\text{lca}(x,y)$-operator. In Algorithm 1, $N_{\text{lca}}$ depends on the closure index $c$ that determines how many times the while loop is done. As this number $c = O(|V|)$ (Lemma 3.2.1), the number $N_{\text{lca}}$ of calls of the $\text{lca}(x,y)$-operator for this algorithm is:

$$N_{\text{lca}}(\text{algo 1}) = \sum_{k=0}^{c} s_k^2 = \sum_{k=0}^{O(|V|)} O(|V|^2) = O(|V|^3)$$

**Algorithm 1. Straightforward $U_{\text{lca}}$-closure algorithm**

Name: Straightforward $U_{\text{lca}}$-closure

Input: a dag $G$ and a set of nodes $S$ of $G$.

Result: $\tilde{S}$ the $U_{\text{lca}}$ closure of $S$.

```plaintext
S_{\text{tmp}} \leftarrow S;

\textbf{do}

\quad S_k \leftarrow S_{\text{tmp}};

\quad S_{\text{tmp}} \leftarrow \emptyset;

\quad \textbf{for each} (x,y) \in S_k \times S_k

\quad \quad S_{\text{tmp}} \leftarrow S_{\text{tmp}} \cup \text{lca}(x,y);

\textbf{end}

\textbf{while} S_k \neq S_{\text{tmp}}

\textbf{return} S_k
```

**Algorithm 2. Slightly improved $U_{\text{lca}}$ closure algorithm**

Name: Slightly improved $U_{\text{lca}}$-closure

Input: a dag $G$ and a set of nodes $S$ of $G$.

Result: $\tilde{S}$ the $U_{\text{lca}}$ closure of $S$.

```plaintext
S_k \leftarrow S; S_{\text{new}} \leftarrow S;

\textbf{do}

\quad S_{\text{tmp}} \leftarrow S_k;

\quad \textbf{for each} (x,y) \in S_k \times S_{\text{new}}

\quad \quad S_{\text{tmp}} \leftarrow S_{\text{tmp}} \cup \text{lca}(x,y);

\textbf{end}

\quad S_{\text{new}} \leftarrow S_{\text{tmp}} \cup \text{lca}(x,y);

\textbf{while} S_{\text{new}} \neq \emptyset

\textbf{return} S_k
```

This high complexity is due to the fact that some $\text{lca}(x,y)$ are computed several times, for instance those of $S$ are computed $c$ times. This is obviously useless, and Algorithm 2 computes $\text{lca}(x,y)$ only when needed, i.e. only if $x$ and/or $y$
are/is in $S_{\text{new}}$. Yet the $lca$ should be computed for any couple of vertices of $\tilde{S}$. It follows that:

$$N_{lca}(\text{algo2}) = \sum_{x,y \in \tilde{S}} 1 = O(|V|^2) \quad (4)$$

The time complexity of this second algorithm is $O(|V|^2)$ times the complexity of the $lca(x,y)$ subroutine.

Several optimized algorithms are available to retrieve one vertex $v \in lca(x,y)$. We stress the word “one” because, in our approach we need all the vertices of $lca(x,y)$. Most optimized solutions to retrieve one vertex $v \in lca(x,y)$ rely on the shortest path algorithm. After preprocessing, done in $(|V|^{\omega})$ with $\omega = 2.688$, a representative $lca$ can be obtained in constant time for any couple of nodes (see [30-32]). An improved solution, with $\omega = 2.575$, is described in [34]. As the domain is active, [35] have given an algorithm that calculates all $lca(x,y)$ for all pairs of vertices with a mean time complexity $O(|V|^{3.3399})$ and worst time complexity $O(|V|^{3.3399})$.

Recall that we need the $U_{lca} - \text{closure}$ of a set $S$ (denoted by $\tilde{S}$) and Algorithm 2, even with the optimized subroutine [35] for computing $lca$, would have a complexity of at least $O(|V|^{3.3399})$ for computation time and $O(|V|^2)$ for memory space (in order to store pre-computed $lca$). We thus introduce an optimized dedicated solution that takes the topological vertex order induced by the $\text{dag}$ into account. This solution has lower worst time complexity – $O(|\tilde{S}| |E|)$ – and space – $O(|\tilde{S}| |V|)$ – complexity. The advantage of this solution is even more relevant in practice, since for most real cases $|\tilde{S}| \ll |V|$ and for most ontologies $|E| \ll |V|^2$.

### 4.2 Optimized algorithm to compute $U_{lca}$ closure in $O(|\tilde{S}| |E|)$

This subsection details an optimized algorithm that determines the $U_{lca}$ closure of $S$ for a $\text{dag}$ $G = (V,E)$ in $O(|\tilde{S}| |E|)$. The key idea of this algorithm is that, although there are $O(|V|^2)$ couples of vertices, at most $O(|V|)$ nodes can be added to $S$. Rather than computing the $lca$ for each pair of vertices, we thus consider each node and decide greedily whether or not it must be added to $S$. This can be done efficiently by taking the topological vertex order induced by the $\text{dag}$ into account. These are classical algorithms that we recall below so that the paper will be self-contained.

#### Algorithm 3. Post order implementation (recursive and iterative subroutines).

```plaintext
Name: postOrderRec
Input: a dag $G$, a node $n$ of $G$
Result: add the list of desc($n$) in post order to the postOrder list of $G$
mark $n$ as visited
for $s$ in children($n$)
  if $s$ has not been visited
    postOrderRec($s$)
end
append $n$ to the postOrder list of $G$
```

Our $U_{lca}$ closure algorithm considers vertices in post order, i.e. a vertex is never considered before considering all of its descendants. Indeed, vertices of a $\text{dag}$ can be ordered along a horizontal line such that all descendants of a vertex are placed to its left. We call this a post order since, as shown in [36], one can be efficiently obtained using the post-order indices of a depth-first search (Algorithm 3).

#### Algorithm 4. Computation of $U_{lca}$ closure

```plaintext
Name: U_{lca} closure
Input: a dag $G$, a set $S$ containing nodes of interest.
Result: the $U_{lca}$ closure of $S$. 
$S^{AL} \leftarrow \emptyset$
P $=$ postOrder($G$)
for $n$ in $P$
  $S^D(n) \leftarrow \emptyset$ // $S^D(n)$ is the set of descendants of $n$ present in the $U_{lca}$ closure of $S$
  max$S^D(n)$ $\leftarrow 0$ // max$S^D(n)$ is the maximal value $\text{max}(S^D(s))$ with $s$ a child of $n$
  for $s$ in children($n$)
    $S^D(n) \leftarrow S^D(n) \cup S^D(s)$
    max$S^D(n)$ $\leftarrow \text{max}(S^D(s)$ \& $\text{max}$ $S^D(n))$ (*)
  end
end
$S^{AL} \leftarrow S^{AL} \cup \{n\}$ (**)
return $S^{AL}$
```

#### Proposition 4.2.1. (Proof of correctness). Given the inputs $G = (V,E)$ and $S$, the set $S^{AL}$ returned by Algorithm 4 is the closure of $S$ with respect to $G$, $S^{AL} = \tilde{S} = \tilde{S}$.

#### Proof. Let $P$ denote the array of nodes of $G$ sorted by the postOrder function. The $U_{lca}$ closure algorithm goes through $P[1], ..., P[k], ..., P[|V|]$ gathering, for each $k$, a subset of $S^{AL}$ denoted $S^{AL}(k)$. It is clear that $S^{AL} = S^{AL}(V)$. We show by induction that:

$$S^{AL}(k) = \tilde{S} \cap \text{desc}(P[k]), k = 1, 2, \ldots, |V|. \quad (5)$$
The statement (5) holds for $k = 1$.
Node $P[1]$ has no descendant. It is kept in $S^{AL}$ if and only if $P[1] \in S \subseteq \tilde{S}$ (the line (***) of the algorithm.) Thus, $S^{AL}(1) = \tilde{S} \cap P[1..1]$ and $S^D(P[1]) = \tilde{S} \cap \text{desc}(P[1])$.

Assuming that (5) holds for indices $1, ..., (k - 1)$, then (1) also holds for indices $1, ..., k$.

Let $n = P[k]$ be the current node and $\{s_1, ..., s_p\}$ be the set of immediate descendants (children) of $n$. Since nodes are considered in post order, all children of $n$ are at the left of $n$ in array $P$. When the algorithm is considering $n$, all of its children $s_1, ..., s_p$ have already been treated and set $S^D(s_i)$ has been recorded for each $s_i \in \{s_1, ..., s_p\}$. At the point (**), the current recorded set for node $n$ (see point (**)) is:

$$S^D(n) = S^D(s_1) \cup ... \cup S^D(s_p) \quad (6)$$

The test at the point (***) of the algorithm is used to decide whether or not $n$ is in the closure and should be added to $S^D(n)$ and to $S^{AL}$.

- If $n \in S$, $n$ is added to $S^{AL}(k)$ as well as to $S^D(n)$ and evidently $n \in \tilde{S} \cap P[1..k]$ and $S^D(P[k]) = \tilde{S} \cap \text{desc}(P[k])$.
- If $n \notin S$. In this case, by the induction hypothesis, at the point (***) any node of $S^D = S^D(s_1) \cup ... \cup S^D(s_p)$ are in $\tilde{S}$ and we have $S^D = S^D(n) - \{n\} = \tilde{S} \cap \text{desc}(n)$. The only thing remaining to prove is that $n$ will be included in $S^{AL}(k)$ and in $S^D(n)$ iff it is the least common ancestor of two nodes of $S^D$.

- If $(S^D(n) \geq \max S^D(n))$ then there are at least two nodes $z, t$ of $S^D$, such that $n = \text{lca}(z, t)$ and $n$ should be added to $S^{AL}(k)$ as well as to $S^D(n)$, (see Fig. 5 for an example) As $(S^D(n) \geq \max S^D(n))$, there are at least two nodes $z, t \in S^D$ such that $\{z, t\} \notin S^D(s_i)$, $i = 1, ..., p$. It follows that there are two distinct children $s_i, s_j$ of $n$ such that $z \in S^D(s_i)$, $t \in S^D(s_j)$. By definition of the $\text{lca}$, $n \in \text{lca}(z, t)$ if and only if, $n \in A(z, t)$ and $n$ has no descendant in the ancestor set $A(z) \cap A(t)$. The former assertion is obvious, let us prove the latter by supposing this is not the case (reductio ad absurdum). So, there is a node $n' \in A(z) \cap A(t)$ and a $(n, n')$ directed path in $G$. This path necessarily goes through a child $s_m$ of $n$ and, by the induction hypothesis, $S^D(s_m) = \tilde{S} \cap \text{desc}(s_m)$. It follows that $(z, t) \notin S^D(s_m)$, which is impossible.

- If the test (***) is not true, then $n$ is not in the closure and is added neither to $S^D$ nor to $S^{AL}$.

The main thing to prove is that when $(S^D(n) = \max S^D(n))$, there are not two nodes $z, t$ of $S^D(n)$ such that $n = \text{lca}(z, t)$. (see Fig. 6 for an example). As $(S^D(n) = \max S^D(n)$ then there is some $i \in \{1, ..., p\}$ such that $S^D(n) = S^D(s_i)$. In this case, $n$ cannot be the $\text{lca}$ of a couple of nodes $(z, t)$ because node $s_i$ is (by construction) an ancestor of $z, t$ and a descendant of $n$. It follows that $n \notin \tilde{S}$ and the proof is complete. □

**Proposition 4.2.2.** (Time complexity of $U_{\text{lca}}$ closure algorithm) For a node set $S$ in a dag $G = (V, E)$, the $U_{\text{lca}}$ closure algorithm runs in $O(\langle S \rangle |E|) = O(|V||E|)$.

**Proof.** Obtaining the postOrder vector of nodes is done through a classical depth first search traversal of the graph in $O(|E|)$. The complexity of the remaining part of the algorithm, made of two nested for loops, is obviously determined by the number of executions of line (**). This line computes the union of two sets of at most $|S|$ elements and is executed for every child of every node, i.e. $O(|E|)$ times. It follows that the whole complexity of this algorithm is $O(\langle S \rangle |E|)$. In the worst case, $|S|$ equals $|V|$ leading to a complexity of $O(|V|^2 |E|)$. Note however that generally $|S| \ll |V|$.[.] □

**Proposition 4.2.3.** (Space complexity of $U_{\text{lca}}$ closure algorithm) For a node set $S$ in a dag $G = (V, E)$, the $U_{\text{lca}}$ clo-
An efficient algorithm has been described to compute the $|S|\times|V|$ = $O(|V|^2)$ memory space.

**Proof.** For each node of the outer “for” loop of algorithm 4, a subset of $S$ is stored. In the worst case, $|S| = |V|$ leading to a complexity of $O(|V|^2)$.

Note that in most cases $|S| \ll |V|$. Moreover, when all parents of a node $n$ have been treated, the subset of $S$ attached to $n$ becomes useless, which means that some memory space can be freed. This can be easily done by maintaining a counter for each node initialized to its number of parents. When treating a node, the counters of all of its children are decreased by one, and when a child reaches a zero value its memory is freed. This does not reduce the worst case complexity, since this optimization is useless when the dag is made of one node that has $|V| - 1$ children, but it significantly reduces the memory space needed in most real cases.

The algorithms presented in this section are used for identifying a set of relevant concepts according to those of interest. Their relationships have to be rebuilt in order to obtain the intended sub-ontology.

An efficient algorithm has been described to compute the $U_{lca}$ closure of a set of concepts. Yet, when considering a set of vertices $S$, every greatest common descendant ($U_{gcd}$) of a pair of vertices of $S$ is also a key vertex to grasp relationships between vertices of $S$. All results and algorithms of the previous section concerning $lca$ can easily be transposed to $gcd$. This is due to the duality of the two operators. Let $G$ be the dag $(V, E)$ and let $G^*$ the dag $G^* = (V, E^*)$ such that $(v, u) \in E^*$ iff $(u, v) \in E$. Then, for all $x$ and $y$ of $V$, an $lca$ (resp. $gcd$) of $x$ and $y$ in $G$ is a $gcd$ (resp. $lca$) of $x$ and $y$ in $G^*$. $S$ denotes the $U_{lca}$ closure of $S$ and, similarly, $\overline{S}$ denotes its $U_{gcd}$ closure.

### 5 Building the Sub-ontology

The previous section describes efficient algorithms to compute $U_{lca}$ and $U_{gcd}$ closures of a set of concepts. Both algorithms have been used within a broader process to extract the sub-ontology.

#### 5.1 Using $U_{lca}$ and $U_{gcd}$ closures to define the set of relevant concepts.

Given a set $S$ containing concepts of interest, we define the set $S_r$ of relevant concepts kept in the sub-ontology as $S_r = S \cup S_{lca}$. Note that this set is not closed with respect to $lca$ or $gcd$ operators, as illustrated by Fig. 7, whereas $S_r$ does not. One could thus be tempted to use the set $\overline{S}$ defined by the closure of both operators.

**Definition 5.1.1.** Let the increasing set sequence defined by: $S_0 = S$ and $S_{k+1} = S_k \cup S_{lca}$, $k = 0, 1, ...$

Set $\overline{S}$ is equal to set $S_r$ associated with index $c$ such that $\forall k \geq c$, $S_k = \overline{S} \cup S_{lca}$.

Using this constructive definition, set $\overline{S}$ can be obtained by iterating the $U_{lca}$ closure and $U_{gcd}$ closure algorithms until a fixed point (set) is obtained. This can be done in an $O\left(|S| \times |E|\right)$ algorithm. Yet the concepts of $\overline{S}$ that are not in $S \cup S_{lca}$ are obtained by applying both lca and gcd operators.

For instance, concept $A1$ in Fig. 7 is present in $\overline{S}$ because $A1 = \overline{lca}(lca(C1, C2), gcd(C2, C3))$. The relationship between such concepts and the initial ones is thus hard to interpret and including them in the sub-ontology can be more confusing than helpful. We thus build the sub-ontology with
respect to $S$ using only concepts within $\overline{S} \cup S$ since this is faster and more meaningful than using $\overline{S}$.

Note that the set of concepts derived from the SCA and ASA strategies from Nebot and Berlanga [20], is \{$C_1, C_2, C_3, B, X, A_1, \text{Thing}$\} $\supseteq \overline{S}$. Indeed, though poorly informative, $X$ and $A_1$ are kept because they are common ancestors of $C_1$ and $C_2$. $X$ is discarded by their third strategy (ASA-ST) while the $A_1$ concept may be captured or not, depending on their choice of spanning tree.

### 5.2 Adding hyponymic and hypernymic relationships among relevant concepts

Once the set of relevant concepts $S_r$ has been found, hypernymic (and hyponymic) relationships among them have to be identified. We want to preserve the hypernymic and hyponymic relationships among these concepts of interest (ASA-ST) while the $A_1$ concept may be captured or not, depending on their choice of spanning tree.

**Definition 5.2.1.** Given the dag $G = (V, E)$, the relevant $\text{sub\text{dag}} G_r = (V_r, E_r)$ is defined with:

- $V_r = \overline{S} \cup S$
- $(u, v) \in E_r$ if $u$ is a descendant of $v$ that is present in $V_r$ but not in $\overline{S}$.

The set $E_r$ of edges can be efficiently computed due to the topological order induced by the dag $G$. This time we will consider a vertex $u$ only after having considered all of its ancestors, i.e., the set of Relevant Reachable Ancestors of $u$ containing vertices that are present in $V_r$ and can be reached from $u$ through a path crossing no nodes of $V_r$. When considering vertices in preorder, the set $V_{\text{RRA}}(u)$ of the current node $u$ is simply the union of $V_{\text{RRA}}(f)$ sets of all of its parent nodes that are not in $V_r$, plus all its parent nodes that are in $V_r$. The set $E_r$ is then constructed by adding, for each node $u$ of $V_r$, edges $(u, v)$ between $u$ and any node $v$ of $V_{\text{RRA}}(u)$ as detailed in Algorithm 5.

As for Algorithm 4, the key instruction, line (*), computes the union of two sets of at most $|V_r|$ elements and is executed for every parent of every node of the initial dag $G$ i.e. $O(|E|)$ times. The overall complexity of this algorithm is thus $O(|\overline{S} \cup S| |E|) = O(|V| |E|)$.

### 5.3 Adding other kinds of relationships between relevant concepts

While ontologies may include various kinds of concept relationships, most sub-ontology extractions identify relevant concepts only using the is-a relationship. Indeed this relationship is the only one shared by all ontologies and it constitutes their backbone. The key role of the is-a relationship is clearly explicit in the formal definition of the ontology proposed by [37] (p. 244-). Consequently, our method to select relevant concepts relies only on this relationship. However, other relationships may be added among those selected concepts. This section describes the strategy we used within this second phase.

Having the relevant sub-isa-dag $G_r = (V_r, E_r)$ obtained from the original ontology $\sigma$, one should at least enrich the relevant sub-ontology $\sigma_r$ with any relationship $u \text{R} V$ of $\sigma$ concerning two concepts $u, v$ of $V_r$.

Some additional relationships between two concepts $u, v$ of $V_r$ can be deduced from the original ontology even if the relationship $u \text{R} V$ is not explicitly present in $\sigma$. For instance, let $\text{ Isa, Eq}$ and $P \text{of}$ be the is-a, equivalence and part-of relationships, respectively. If we consider the case where $A \text{isa B, A Pof C and A Eq D}$, and $V_r = \{D, C\}$, the relationship $D \text{ Pof C}$ can be added since $A \text{ Pof C and A Eq D}$.

Formalizing the relationship that should be kept is far from obvious and the semantics of the relationships at stake must be properly taken into account. Yet some general rules can be applied for the simplest cases. In particular, for any transitive relationship $R$ (i.e. when $a \text{R} b$ and $b \text{R} c$ then $a \text{R} c$), we can add the relation $u \text{ R} V$ and $v \text{ R} V$ for any pairs of concepts such that there is a sequence of relationships $a_1 \text{R} a_2,..., a_n \text{R} V$ in $\sigma$ and no $a_i$ concept belongs to $V_r$. This is a straightforward generalization of the approach used to determine the is-a relationship of $\sigma_r$.

Algorithm 5, formulated to add the is-a relationship to $\sigma_r$, relies on the fact that the underlying graph representation of these relationships is a dag. Hence, this algorithm can only handle anti-symmetric relationships (that induce a dag structure) and is only meaningful for transitive relationships (otherwise it will induce wrong relationships). It can thus be used on is-a and part-of relationships but not for the equivalence one that is transitive but not anti-symmetric and may thus induce cycle ($A \text{ Eq B, B Eq C and C Eq A}$).

All of the presented algorithms have been implemented in an application called OntoFocus using Java and the Jena library. OntoFocus takes an OWL ontology and a set of concepts of interest $S$ as input, and an OWL sub-ontology focused on $S$ as output. The current version adds any direct relationships existing between two relevant concepts and uses Algorithm 5 to add direct and induced is-a and part-of relationships between them. Applications are discussed in the following section.
6 Applications, Results and Perspectives

The accumulation of digital information has become problematic in many domains, especially in Life Sciences where genome decryption, for example, leads to hundreds of databases and interacting applications. They often make use of ontologies to organize the information, to support semantic information retrieval and visual navigation through it. The size and scope of bio-ontologies have increased almost as rapidly as the biological data. However, in numerous cases, only a few of their concepts are useful at once. For example, molecular biologists often visualize and analyze a graph representation of the ontology relations of a subset of Gene Ontology concepts, but unfortunately this analysis is often done by hand [38-40].

OntoFocus has been used to restrict the Gene Ontology to the subset of concepts related to the BRCA1 gene associated with Breast Cancer susceptibility. More precisely, we downloaded the annotation of the Swiss-Prot protein P38398 (associated with BRCA1) via the Go Annotation (GOA) service provided by the European Bioinformatics Institute (http://www.ebi.ac.uk/GOA/). We thus obtained 102 redundant annotations corresponding to a set of 50 individual GO concepts. The corresponding sub-ontology inferred by OntoFocus in about one minute contains 92 relevant concepts (instead of thirty thousand).

As illustrated in Fig. 8, the visualization is very comfortable and within human cognitive and perceptive limits. Here two of the three connected parts of the sub-ontology are presented: concepts that correspond to molecular functions (GO_0003674) and those that correspond to cellular components (GO_0005575). The last part corresponding to biological processes is not presented here. However, it contains 63 concepts, which also facilitates visualization.

Giving a specific view on the ontology may highlight part of it and clarify the fact that some concepts are refinements of others. Fig. 8 highlights, for example, that several annotations are refinements of the protein binding function (GO_0005515). Such user-centered sub-ontologies may be useful for biological users exploiting annotations, especially for collaboratively updating gene annotations. This approach is also widespread in comparative genomic analysis in which gene function variation among species (gene annotations, in this case, are species dependant) is studied. An intuitive way to summarize this variation is to color concepts of an ontology excerpt. Such representations have been achieved on a large scale for genes related to human diseases [41] and made available on a Web site (http://www.informatics.jax.org/GOgraphs/OrthoDisease/). These excerpts have been obtained by keeping all ancestor concepts (hence including useless concepts while ignoring common descendants). This rough strategy can thus be advantageously replaced by the one presented in this paper and integrated in OntoFocus.

The same approach may be used to simultaneously apprehend annotations of several genes that share some biological characteristics (e.g. genes having similar expression profiles in microarray experiments.) In such cases, GO concepts are useful for giving biological meaning to the corresponding gene clusters. These annotating genes can be used to automatically define the set of concepts of interest that can serve as an OntoFocus input. The resulting sub-ontology highlights gene annotation relationships and helps to determine the cluster semantic. The additional concepts proposed by OntoFocus are also good candidates for annotating a cluster as a whole. Indeed, as underlined by [26], least common ancestor concepts are of particular interest for summarizing annotations.

Fig. 8. Visualization of molecular function (GO_0003674) and cellular component (GO_0005575). GO-sub-ontologies constructed by OntoFocus using BRCA1 annotation (blue colored concepts were added by OntoFocus to make the sub-ontology explicit).

Our approach, illustrated here by a biological application, may be used in every Information System dealing with amounts of data and huge ontologies. Indeed, having sub-
ontologies may be helpful in any computer assisted ontology operation requiring a human expert (e.g. ontology design and evolution or visual filtering within conceptual maps).

To be effective, visual representations must be semantically guided so that the user may observe semantics while browsing data. Therefore, we plan to adapt sub-ontology extraction in order to enrich the sub-ontology with additional information such as the number of children/parents in the reference ontology and the original path length of an is-a edge. This additional information will allow us to compute the semantic distance separating two relevant concepts with respect to the reference ontology [42].

Concerning ontology visualizations, as mentioned in [13] many papers have been proposed but they all focus on hierarchy representation. “In general the main drawback is that the semantics entailed by the ontologies is not used fully” [3]. We assume that there is an alternative way of representing ontologies. In [43] a semantic distance has been proposed to evaluate similarity between concepts. This distance, combined with visualization techniques using multi-dimensional scaling, provides a different view of the concept hierarchy. Concerning the prospects of the work presented in this paper, we plan to insert OntoFocus in an environment that we have developed: OBIRS (ontology based information retrieval system) [44]. In response to a query built with concepts of a domain ontology, OBIRS builds semantic maps to display the retrieved documents. These maps may be used for navigating through, analyzing or viewing digital documents annotated by concepts of a given ontology.

7 CONCLUDING REMARKS

After justifying the crucial need for sub-ontology extraction methods for different purposes, this paper introduces a graph based definition of relevant concepts for highlighting relationships among concepts of interest and provides an optimized algorithm to identify them. According to a reference ontology and a set of concepts of interest, it computes $U_{lea}$ and $U_{gcd}$ closures on direct acyclic graphs (dag) induced by is-a relationships. It then infers relationships among these relevant concepts to provide a user-centered sub-ontology. The complexity of this algorithm is proportional to the number of kept concepts in addition to the number of reference dag edges.

The resulting OntoFocus program may be freely used online (http://www.ontotoolkit.mines-ales.fr/). Many applications may benefit from this algorithm. One of them, developed in this paper, concerns the life science domain and the use of gene ontology to analyze gene annotations. Because of the very short calculation time (few seconds for about 30,000 concepts), calculations may be performed each time a user needs a specific view. This ensures that the sub-ontology is always compliant with the reference ontology and offers novel prospects for real-time user interactions. For instance, this provides the possibility of chaining several extractions, progressively refining the ontology so that the user can precisely tune the filtering process.

Future directions of our work include ontology visualization, organization of query results and dedicated applications coupled with Gene Ontology.

ACKNOWLEDGMENTS

We would like to thank Pierre Jean for his help in deploying the OntoFocus Web server.

The authors wish to thank the reviewers for their many detailed comments and proposals to improve the paper. This publication is the result of a collaboration between the “Institut des Sciences de l’Evolution de Montpellier” (UMR 5554 - CNRS) and the “LGZP Research Centre from Ecole des Mines d’Alès”. This work was supported by the French Agence Nationale de la Recherche ‘Domaines Emergents’ [ANR-08-EMER-011 ‘PhylAriane’] and is the contribution N°2011-067 of ISEM.

REFERENCES

sy, 2002.


Vincent Ranwez is a research member of the ISEM laboratory team at the University Montpellier II (France). He holds a PhD in Computer Science from the University of Montpellier (France). He previously worked for Clinigetics SA, a biopharmaceutical firm that used DNA chip analyses to identify potential gene targets for new cardiovascular drugs. This work extensively relied on Gene Ontology based annotation. His main research now focuses on molecular evolution, especially phylogeny and supertree inferences. Since phylogenies are represented as trees or as DAG (in case of hybridizations) most of his work is related to graph algorithms.

Sylvie Ranwez is a research member of the LGI2P Research Center team at the Ecole des Mines d’Alès (France). She holds a PhD in Computer Science from the University of Montpellier (France), dealing with automatic composition of adaptive hypermedia documents based on ontologies and intentional user requests. Her research focuses on technical assistance in indexation, navigation and information search using semantic support (ontologies) and visualization. She has published research papers on semantic distance, conceptual maps, ontological and FCA based indexing and visual navigation.

Stefan Janaqi is a research member of the LGI2P Research Center team of the Ecole des Mines d’Alès (France). He holds a PhD in Computer Science from University Joseph Fourier, Grenoble (France), dealing with geometric properties of graphs. His research focuses on mathematical models for optimization, image treatment, evolutionary algorithms and convexity in discrete structures such as graphs. He has published research papers in differential evolution and matching on vision. He applies optimization models to blending in the petroleum industry.