Taxonomy Structure Extraction: The SIFT Algorithm Approach

Jorge Martinez-Gil,

Software Competence Center Hagenberg (Austria), jorgemar@acm.org

Keywords: Algorithms; Knowledge Engineering; Knowledge Integration; Taxonomy Analysis

Abstract

In this paper, we introduce SIFT, a three-phase algorithm designed for analyzing the structural data encapsulated within taxonomies. SIFT's primary strength lies in its ability to harness the inherent hierarchical information of taxonomies, enabling it to deduce relationships that facilitate subsequent merging processes. This approach is especially significant in contexts where traditional taxonomy alignment methods, which rely on text-based information from taxonomy nodes, are ineffective.

1 Introduction

The problem of aligning taxonomies is one of the most interesting and relevant issues for knowledge engineers, since it has implications in a wide range of computational problems including file system merging, creation of operating systems distributions, catalog integration, distributed taxonomy search, and so on. The non-deterministic nature of the problem is given by the fact that not even humans are able to identify optimal alignments [6], so the process is highly subjective. This means that its boundaries often go further than the category of an engineering problem what makes difficult to find closed solutions in this area. However, the large amount of people, research groups and resources dedicated to provide solutions in this field, tells us that we are facing a key challenge in order for a convincing way to automatically integrate taxonomic knowledge to become real.

In the last years, the need for methods to integrate knowledge has increased. Note that the need for aligning taxonomies comes from the old field of database schema integration [13]. This field was born to work in a unified way with databases which had been developed independently. Nowadays researchers aim to make the techniques for aligning knowledge models flexible and powerful enough to work with all kind of database schemas, XML schemas, taxonomies, E/R models, dictionaries, and

so on. Therefore, the problem we are facing consists of providing a set of correspondences between the nodes of two taxonomies about the same domain but which have been developed separately [22].

The major contribution of this work is the proposal of a 3-step algorithm that is able to analyze the structural information represented by means of a taxonomy. The major advantage of this analysis is that it can allow us to leverage the information inherent to the hierarchical structure of taxonomies to infer correspondences which can allow to automatically merging them in a later step. This is particular relevant in scenarios where taxonomy matching techniques exploiting textual information from the taxonomy nodes cannot operate successfully.

From now on, this work is structured in the following way: The second section describes the state-of-the-art on taxonomy alignment. The third section describes the design and development of the algorithm. Case studies section provides some scenarios where our algorithm can help to solve real problems, including a brief discussion on the strengths and weaknesses of the proposal. Finally, we outline the key points of our contribution and propose future research tasks.

2 Related Work

The problem of aligning taxonomies have received much attention by the research community since various knowledge based applications, including clustering algorithms, browsing support interfaces, and recommendation systems, perform more effectively when they are supported with domain describing taxonomies, which help to resolve ambiguities and provide context [3]. Furthermore, this problem is of great interest on a number of application areas, especially in scientific [6], business [1] [17], and web data integration [4] [20].

Taxonomy alignment techniques are able to detect taxonomy concepts that are equivalent. But, when can we say that two concepts are equivalent? If we attend only to the text label for representing the concepts, we can find many examples in everyday life, for instance, *lift* and *elevator* or *car* and *automobile* seem to be equivalent concepts since they represent the same real idea or object [11]. However, it is well known that when taxonomies are used as knowledge sources, the way users perceive the degree of likeness between pairs of concepts is highly dependent on the domain being explored [3]. Therefore, synonymy between text labels is not always an equivalence indicator, so it is necessary to focus in the context the concepts are being considered [15].

Existing taxonomy alignment techniques focus on different dimensions of the problem, including whether data instances are used for matching [12], whether linguistic information and other auxiliary information are available [16], and whether the match is performed for complex structures [19]. Our algorithm fits in this last category.

Algorithms implementing techniques for matching complex structures are mostly based on heuristics. Heuristics consider, for example, that elements of two distinct taxonomies are similar if their direct sub-concepts, and/or their direct super-concepts and/or their brother concepts are similar [21]. These structural techniques can be based on a fixed point like that proposed in [8], or can be viewed as a satisfiability problem of a set of propositional formulas [2]. There are also some proposals to align taxonomies supposed to be asymmetric from a structural point of view [5], or to create matching functions by means of a composition of various techniques designed to make best use of the characteristics of the taxonomies [21].

Despite such advances in matching technologies, taxonomy alignments using linguistic information and other auxiliary information are rarely perfect [9]. In particular, imperfection can be due to homonyms (i.e., nodes with identical concept-names, but possibly different semantics) and synonyms (concepts with different names but same semantics [14]). However, the major advantage of pure structural matching techniques is that finding perfect alignments is possible in many cases [10].

3 Contribution

We approach the problem from the classic perspective, that it is to say, a taxonomy can be defined as a set of concepts that have been hierarchically organized to control the terms belonging to a vocabulary. The goal is to facilitate a number of operations on items from a repository. However, a problem occurs when two item repositories have to be merged, since it is also necessary to merge the two taxonomies which describe them.

Our contribution to face this problem is the proposal of an efficient 3-step algorithm for the analysis of taxonomies describing such repositories. This analysis could be helpful for solving the problem of heterogeneity between the given taxonomies from a strictly structural point of view in a later step. As a collateral effect, the output data from our algorithm could be also used for exploiting any kind of solution involving the use of information from the structure of the given taxonomies. Use cases

section will explore this in more detail.

More formally, we can define a mapping as an expression that can be written in the form *(c, c', n, R)*. Where *c* and *c'* are concepts belonging to different taxonomies, *R* is the relation of correspondence and *n* is a real number between 0 and 1. *n* represents the degree of confidence for *R*. In our work, *c* and *c'* will be concepts represented by means of taxonomy nodes (a.k.a. taxons) which are assigned a rank and can be placed at a particular level in a systematic hierarchy reflecting relationships. Moreover, the relation *R* which describe how *c* and *c'* are related is going to be of similarity.

The algorithm that we propose is divided into three high level steps. The first step is optional since it is only necessary when the given knowledge model is not a taxonomy yet, but another kind of more general model like an graph or an ontology [23].

1. To convert the knowledge model into a taxonomy (See Algorithm 1).
2. To store the taxonomy in some parts of a special data structure (See Algorithm 2).
3. To order and fill the data structure with complementary calculations (See Algorithm 3).

Finally, it is necessary to call the algorithm (See Algorithm 4). The philosophy of the algorithm consists of detecting the changes in the depths of each taxon in the hierarchy. In this way, it is possible to count the different kinds of neighbors that a concept may have.

Before designing the algorithm, it is also necessary to define a data structure (DS) to store the data calculated by the algorithm. The data structure is a linked list with six records in each node: depth, children, brothers, brothers_left, same_level and name. Table 1 tells us the data type and a brief description of each of these records. In the next subsections, we are going to describe more in depth each of the main steps of the proposed algorithm.

3.1 Converting a knowledge model into a taxonomy

This is the first step which consists of converting the model into a taxonomy which will allow us to compute more easily the data related to the neighborhood of each concept into the knowledge model. This step is optional and it is only necessary when the input is not a perfect hierarchy but contains some cycles. This is the usual case when working with graph models or ontologies, for example. The

Attribute	Type	Description
depth	integer	Level of the current taxon (begins with 0)
children	integer	Number of children of the current taxon
brothers	integer	Number of brothers of the current taxon
brothersLeft	integer	Number of brother taxons that are above this
sameLevel	integer	Number of taxons with the same depth
name	string	ID of the taxon

Table 1: A node of the linked list which stores the information

procedure is inspired by one proposed in [18] to visit all the concepts in an ontology. Algorithm 1 shows the related portion of pseudocode.

Algorithm 1 ont2tax: Procedure for converting a generic knowledge model into a taxonomy

Require: cls: *class*, occurs: *list*, depth: *integer*

```
1:  storingInTax(cls, depth); Step 2
2:  if (cls.canAs( model.class ) AND (NOT occurs.contains( cls ))) then
3:     while iterator = cls.SubClasses do
4:        class sub := (class) iterator.next
5:        occurs.add(cls)
6:        ont2tax (sub, occurs, depth + 1)
7:        occurs.remove(cls)
8:     end while
9:  end if
10: return true
```

3.2 Storing the taxonomy in the data structure

In this second step, we only know the depth (number of indents for the taxon) and the name of each concept, so we can only partially fill the data structure, thus, we can only invoke the procedure with the arguments depth and concept name.

Algorithm 2 storingInTax: Storing the taxonomy in the data structure

Require: cls: *ontology*, depth: *integer*

```
1: Element e := new Element (depth, 0, 0, 0, 0, cls.getName)
2: DS.add (e)
3: return true
```

3.3 Ordering and filling the data structure

With data stored in the DS, we can now detect the changes in the depth of the entries in the taxonomy to compute the number of children, brothers and, so on. It is necessary to take into account the following rules:

1. All taxons with the same depth are same level taxons.
2. A chain of brothers is a chain of taxons at the same level.
3. A change to an outer taxon breaks a chain of brothers.
4. All brothers with a previous position are on the left.
5. Given a taxon, if the following concept has an inner depth, it is a child.
6. A chain of children can only be broken by a change to an outer taxon.
7. An inner taxon (grandson taxon) does not break a chain of children.

Algorithm 3 shows us the procedural implementation for this set of rules. The computational complexity of this procedure is low, even in the worst of cases we would have $O(n^2)$, since the most complex portion of code can be implemented by means of two simple loops. This means that our solution presents a great scalability regardless of the platform on which the algorithm could be implemented and executed.

3.4 Calling to the algorithm

Now, it is necessary to invoke the algorithm. At this point it is necessary to define the taxonomy model and to locate the concepts without ancestors, in order to begin to visit all the concepts. This is particular relevant in forest models[1]. Note that the ArrayList is necessary to store the visited concepts. Algorithm 4 shows the related portion of pseudocode.

[1]Forest model is that kind of graph model where there is no connection between some graph components

Algorithm 3 finalStep: Ordering and filling the data structure

```
Require: children, brothers, brothers left: integer
Require: same level, i, j, k, t: integer
Require: ag: boolean
 1: for i := 0 to DS.size do
 2:     children, brothers, brothers left := 0
 3:     for j := 0 to DS.size do
 4:         if if (j < i) then
 5:             if if (DS[i].depth = DS[j].depth) then
 6:                 brothers++
 7:                 brothers left++
 8:             end if
 9:             if (DS[i].depth < DS[j].depth) then
10:                 brothers := 0
11:                 brothers left := 0
12:             end if
13:         end if
14:         if (j > i) then
15:             if (DS[i].depth = DS[j].depth) then
16:                 brothers++
17:             end if
18:             if (DS[i].depth < DS[j].depth) then
19:                 break
20:             end if
21:         end if
22:         if  ((j = i+1) AND (DS[i].depth = DS[j].depth - 1) AND (NOT ag)) then
23:             for for k := j to DS[j].depth < DS[k].depth do
24:                 if (DS[j].depth = DS[k].depth) then
25:                     child++
26:                     ag := true
27:                 end if
28:             end for
29:         end if
30:     end for
31:     for for t := 0 to DS.size do
32:         if if (NOT t=i) AND (DS[i].depth = DS[t].depth) then
33:             same level++
34:         end if
35:     end for
36:     DS[i].addNumChildren (children)
37:     DS[i].addNumBrothers (brothers)
38:     DS[i].addNumBrothersOnTheLeft (brother left)
39:     DS[i].addNumSameLevel (same level)
40: end for
41: return  true
```

Algorithm 4 calling to the 3-step algorithm

```
1: Model m := createModel
2: Iterator i := m.listHierarchyRootClasses()
3: while i.hasNext() do
4:     onto2tax((Class) i.next(), new ArrayList(), 0)
5: end while
6: finalStep ()
```

4 Case studies

The purpose of this section is to show the relative ease with which a taxonomy analysis can be performed or a new taxonomy matcher can be developed, based on the data obtained from the algorithm. In the next subsections we are going to show three use cases: how to use the algorithm to compute the leaves in a taxonomy, how to use it to obtain the structural index of a taxonomy, and finally how to use it to align taxonomies automatically.

4.1 Computing the number of leaves in a taxonomy

There are techniques that compute the leaves in a graph for performing a graph analysis. In this sense, our algorithm is easy to extend in order to compute the number of leaves in a taxonomy. To do so, it is only necessary to compute the number of the deepest taxons. We are going to see how to compute the leaves of the taxonomy for an example but, it is possible to compute other features such as paths. Algorithm 5 shows us how to compute the leaves (i.e. terminal nodes) of a given taxonomy.

Algorithm 5 leaves: computing the leaves of a taxonomy

```
Require: var max, leaves: integer
 1: max := leaves := 0
 2: for i := 0 to DS.size do
 3:     if (DS[i].depth > max) then
 4:         max := DS[i].depth
 5:     end if
 6: end for
 7: for for j := 0 to DS.size do
 8:     if (DS[j].depth = max) then
 9:         leaves++
10:     end if
11: end for
12: return leaves
```

4.2 Comparing structural similarities

It is possible to use our algorithm for extracting structural indexes of taxonomies in order to compare its structural similarity. The structural index of a taxonomy is a kind of hash function that tells global information about the total number of children, brothers and so on.

As we show in the state-of-the-art, some techniques use statistical methods for obtaining the structural similarity. It can be useful for adjusting the quality of the generated mappings, for example.

Algorithm 6 shows how to automatically compute one possible structural index from a taxonomy.

Algorithm 6 structuralIndex: extract a structural index of the ontology

```
Require: var acum : integer
1: acum := 0
2: for i := 0 to DS.size do
3:     acum := acum + DS[i].depth
4:     acum := acum + DS[i].children
5:     acum := acum + DS[i].brothers
6:     acum := acum + DS[i].leftbrothers
7:     acum := acum + DS[i].samelevel
8: end for
9: return acum
```

Obviously, when comparing two structural indexes, the higher percentage, the higher the structural similarity of the compared taxonomies. This means that if two taxonomies share the same structural index, we can state that its structural organization is equivalent.

4.3 Real alignment situations

Our algorithm also allows that information to be obtained from the analysis phase can be helpful in order to take decisions in taxonomy alignment scenarios. Output data from SIFT allow us to easily create customized rule-based matchers to obtain more accurate taxonomy alignments. For example, the similarity between two taxonomy concepts or taxons could be given by certain rules concerning ancestors, brothers, and so on.

Moreover, it is possible to combine our proposal with other basic matching algorithms. This can be done by designing a formula that may allow us to align taxonomies from the point of view of the elements, and from the taxonomy structure. This is possible due to the fact one of the attributes (name) contains information at the element level, so it is possible to exploit this kind of information by

using some kind of computational method like the Levenshtein algorithm [7] which is able to calculate similarity between two text strings. In this way, if many attributes (whether structural or textual) are similar, the concepts are also supposed to be similar.

5 Conclusions & Future Work

In this work, we have designed and implemented, SIFT that is a 3-step algorithm that allows us to analyze the structural information inherent to the hierarchical structures of taxonomies. This can be useful when solving problems concerning heterogeneity between taxonomies describing a same domain but which have been developed separately. Therefore, the algorithm that we propose is valid for taxonomy alignment, but also for aligning ontologies, directory listings, file systems, operating system distributions, and in general whatever kind of model which can be transformed into a taxonomy. Our algorithm tries to leverage the inherent characteristics from taxonomies to infer correspondences which can allow us to merge them in a later step, even without text labels describing each of the nodes from the taxonomy.

As future work, we should work to leverage the good performance of our algorithm by designing a combined alignment strategy. In this work, we have proposed to use each of the attributes with similar weights. However, this strategy could not be optimal in some specific cases. We aim to redefine this strategy so that a preliminary study should try to automatically determine the kind of problem we are facing at a given moment, and dynamically assign higher weights to the most promising taxon attributes.

Source Code

An implementation of this algorithm can be found at `https://github.com/jorgemartinezgil/sift`

References

[1] S. S. Aanen, L. J. Nederstigt, D. Vandic, and F. Frasincar. Schema - an algorithm for automated product taxonomy mapping in e-commerce. In *ESWC*, pages 300–314, 2012.

[2] P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In *International Semantic Web Conference*, pages 67–81, 2005.

[3] K. S. Candan, M. Cataldi, M. L. Sapino, and C. Schifanella. Structure- and extension-informed taxonomy alignment. In *ODBIS*, pages 1–8, 2008.

[4] J. Gracia and E. Mena. Semantic heterogeneity issues on the web. *IEEE Internet Computing*, 16(5):60–67, 2012.

[5] F. Hamdi, B. Safar, N. B. Niraula, and C. Reynaud. Taxomap alignment and refinement modules: results for oaei 2010. In *OM*, 2010.

[6] J. J. Jung. Taxonomy alignment for interoperability between heterogeneous digital libraries. In *ICADL*, pages 274–282, 2006.

[7] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, February 1966.

[8] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In *VLDB*, pages 49–58, 2001.

[9] J. Martinez-Gil. An overview of textual semantic similarity measures based on web intelligence. *Artif. Intell. Rev.*, 42(4):935–943, 2014.

[10] J. Martinez-Gil. Automated knowledge base management: A survey. *Comput. Sci. Rev.*, 18:1–9, 2015.

[11] J. Martinez-Gil, E. Alba, and J. F. Aldana-Montes. Optimizing ontology alignments by using genetic algorithms. In C. Guéret, P. Hitzler, and S. Schlobach, editors, *Proceedings of the First International Workshop on Nature Inspired Reasoning for the Semantic Web, Karlsruhe, Germany, October 27, 2008*, volume 419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[12] J. Martinez-Gil and J. F. Aldana-Montes. Reverse ontology matching. *SIGMOD Record*, 39(4):5–11, 2010.

[13] J. Martinez-Gil and J. F. Aldana-Montes. Evaluation of two heuristic approaches to solve the ontology meta-matching problem. *Knowl. Inf. Syst.*, 26(2):225–247, 2011.

[14] J. Martinez-Gil and J. F. Aldana-Montes. Knoe: A web mining tool to validate previously discovered semantic correspondences. *J. Comput. Sci. Technol.*, 27(6):1222–1232, 2012.

[15] J. Martinez-Gil and J. F. Aldana-Montes. An overview of current ontology meta-matching solutions. *Knowl. Eng. Rev.*, 27(4):393–412, 2012.

[16] J. Martinez-Gil and J. F. Aldana-Montes. Semantic similarity measurement using historical google search patterns. *Information Systems Frontiers*, 15(3):399–410, 2013.

[17] J. Martinez-Gil, I. Navas-Delgado, and J. F. Aldana-Montes. Maf: An ontology matching framework. *J. Univers. Comput. Sci.*, 18(2):194–217, 2012.

[18] B. McBride. Jena: A semantic web toolkit. *IEEE Internet Computing*, 6(6):55–59, 2002.

[19] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE*, pages 117–128, 2002.

[20] S. P. Ponzetto and R. Navigli. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *IJCAI*, pages 2083–2088, 2009.

[21] C. Reynaud and B. Safar. When usual structural alignment techniques don't apply. In *Ontology Matching*, 2006.

[22] P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.

[23] S. Sun, D. Liu, and G. Li. The application of a hierarchical tree method to ontology knowledge engineering. *International Journal of Software Engineering and Knowledge Engineering*, 22(4):571, 2012.