Journal of Universal Computer Science, vol. 9, no. 6 (2003), 520-529
submitted: 27/1/03, accepted: 17/3/03, appeared: 28/6/03

Pruning-based Identification of Domain Ontologies

Raphael Volz
(FZI Research Center for Information Technologies
Karlsruhe, Germany
volz@fzi.de)

Rudi Studer
(FZI Research Center for Information Technologies
Karlsruhe, Germany
studer@fzi.de)

Alexander Maedche
(FZI Research Center for Information Technologies
Karlsruhe, Germany
maedche@fzi.de)

Boris Lauser
(Library + Documentation Systems Division
FAO of the UN
Rome, Italy
boris.lauser@fao.org)

Abstract: We present a novel approach of extracting a domain ontology from large-scale thesauri. Concepts are identified to be relevant for a domain based on their frequent occurrence in domain texts. The approach allows to bootstrap the ontology engineering process from given legacy thesauri and identifies an initial domain ontology that may easily be refined by experts in a later stage. We present a thorough evaluation of the results obtained in building a biosecurity ontology for the UN FAO AOS project.

Key Words: ontologies, pruning, knowledge management, structural computing
Category: H.3.7, H.5.4

1 Introduction

The management of large amounts of information and knowledge is of ever increasing importance in today's large organisations. With the ongoing ease of supplying information online, especially in corporate intranets and knowledge bases, finding the right information becomes an increasingly difficult task. Ontologies have been proposed to be a solution to this problem and have been successfully applied to improve knowledge management [1] and search in specialized domains [9].

However, the task of constructing an ontology still requires much effort and is often carried out in an ad-hoc manner. Only few methodologies exist [2, 10] to improve the latter situation and are often extremely complex requiring extensive training and expertise.

We present a novel approach to acquire an initial application ontology for the management of document collections. Our approach builds on the reuse of existing thesauri. Many companies have elaborated taxonomies of products, services and corporate thesauri to ensure proper use of terminology in internal and external documents. Such resources form an important intellectual asset of the business and maybe reused to form an initial ontology.

However, the utility of the ontology in document management is judged by the quality of document retrieval. This quality is largely influenced by the consonance of ontological terms with keywords occurring in managed documents. Unfortunately, this consonance is not met in large-scale thesauri. Hence, we developed an ontology pruning approach, which removes unneeded terms from the thesaurus via a heuristic analysis of terms contained in a document collection. Thereby we ensure that the ontology is focused to the intended document collection.

The usefulness of ontology pruning is emphasized by the results that could be obtained in the AOS project carried out by UN FAO, where we were challenged to acquire an initial ontology for document management in the biosecurity domain.

The paper is structured as follows. Section 2 details the pruning approach. We then elaborate possibilities for evaluation of the pruning approach in section 3. Section 4 presents the results of the evaluation carried out in the context of the UN FAO AOS project [6]. We conclude summarizing our contribution and discussing possible future directions.

2 Pruning

2.1 Pruning in a nutshell

Pruning presents a completely automatic bootstrapping approach for ontology development. Input to pruning is an already existing vocabulary (light-weight ontology, thesaurus or taxnomy), which constitutes a light-weight conceptualization.

The aim of pruning is to automatically extract the subset of the conceptualization which is relevant to the target domain. Naturally, this assumes and requires that the input conceptualization generalizes the target domain.

The decision on whether or not concepts are relevant to the domain is based on a heuristic analysis of document collections. These heuristics operate on a frequency analysis of words that can be extracted from the documents. However, the extraction of relevant terms is based on two sets of documents, one of which contains domain specific documents and the other generic documents. This ensures that we may consider the relative importance of domain terms (wrt. to generic terms) in the pruning process.

Clearly, the identification of a representative set of documents, that represents the domain of interest and that contains concepts relevant to the domain, is central to our approach. Hence, this domain-specific corpus has to be carefully chosen by subject specialists in the area.

The choice of a generic document corpus is deliberate. Generic reference corpora used in the information retrieval community such as CELEX or public news archives have shown to be well-suited in our experiments. As mentioned before, the generic corpus serves as a reference for comparsion with the domain corpus.

2.2 Pruning heuristics

2.2.1 Computing important concepts

The pruning heuristics are based on a frequency analysis of concepts. Concepts are identified in the text via those words, which are used as their lexicalizations. The computation of frequencies for concepts can build on measurements like simply counting the occurrence of words in documents. The latter measure is known as *Term Frequency (TF)* in the information retrieval community. In our work we also used a more elaborate measurement known as *TF/IDF*[1] [4], which punishes concepts that occur often in many documents. This is achieved by normalizing the term frequency number attaching with a term-weighting factor (IDF). For our purpose, we used a weighting factor introduced in [8] which relates the document frequency (DF) of a concept x with the size of the document set:

$$TFIDF(x) = TF(x) * ln(\frac{|corpus|}{DF(x)})$$

2.2.2 Comparing frequencies

In our approach different comparison strategies can be chosen by the user. First, the user may consider different granularities. The granularity ”ALL“ compares the frequencies regarding all documents in the respective sets. On the opposite end the granularity ”ONE“ would consider a domain concept relevant if it occurs more often in some domain document than in any generic document. Other granularities, e.g. comparing the average frequencies, are of course possible but are not yet evaluated.

Second, users may specify a minimum multiplicity factor r, that specifies how much more frequent a domain concept should be compared to a generic concept. Hence, a concept will be considered domain relevant only, if its weighted term frequency is at least r times higher than its counterpart in the generic corpus.

2.3 Concept Acquisition

The ontology pruner has to identify concepts in a document in order to compute concept frequencies. Concepts are linked with (possibly multiple) lexicalizations. Whenever we can identify such a lexicalization in a document the respective frequency of the concept is incremented. This allows to consider synoyms for concepts which are usually available in thesauri. For example, the English word ”bank“ has at least 10 different senses connoting financial institutions, certain flight maneuvers of aircrafts, etc.

All frequencies obtained for individual concepts are aggregated upwards through the taxonomy to ensure that top-level concepts are properly reflected. Via this aggregation we can ensure that top-level concepts that are usually not frequently used in the texts are not considered as being irrelevant for the target domain.

We evaluated two alternatives for the identification of lexicalizations (cf. Figure 1), one treats documents as a vector of words[2] the other tries to match lexicalizations with

[1] Term Frequency / Inverted Document Frequency

[2] each word is separated from others by whitespace or punctuation

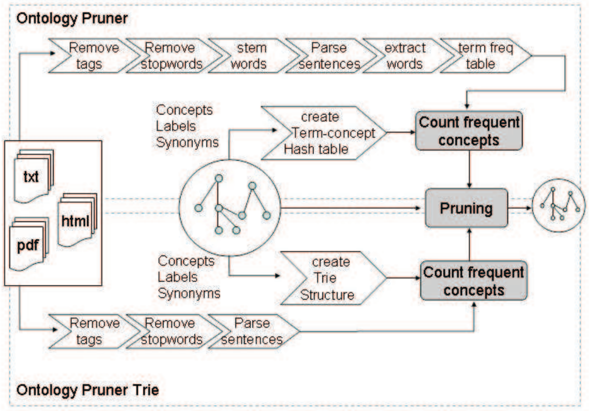


Figure 1: Identification of concepts in text

document content using a TRIE-based structure [3]. The latter can cope with compound terms such as ”food safety“ since whitespace can occur within a TRIE path.

In order to cope with different document formats, e.g. HTML or PDF, the first step in the pruning process removes document specific markup leading to a plain text representation. Then a stop-word list is applied to filter out language specific fill words (such as ’and’, ’in’, etc.). The remaining steps are specific to each method of concept acquisition.

The vector-based pruner uses shallow natural-language processing techniques to stem words and builds up a concept frequency vector. The word vectors of all documents are assembled into one term-frequency table which is used as a basis for further computations. The identification of concepts is accomplished using a term-concept hash table, which allows to look-up concepts via their lexicalizations.

The TRIE-based pruner incrementally processes each pre-processed document by means of a TRIE data structure. It counts the frequency of respective concepts, whenever a leaf of the TRIE is reached (i.e. a label or synonym has been found in the text). In the current implementation, each occurrence of a lexicalization increases the frequency count of a concept by 1, no matter where the lexicalization appears in the text.

3 Evaluation Desiderata

This section discusses how a sensible evaluation of the pruning of a conceptualization can be carried out.

3.1 User Parameters

As discussed in the previous section the ontology pruning process may be influenced by the user using several parameters. The evaluation should clarify the effects of the three main parameters frequency weighting measure (TF, TF/IDF), granularity (One, All) and ratio on the output. Second, we may evaluate the effects of the two approaches on concept identification (Vector, Trie).

3.2 Resource Selection

Naturally, the effects of pruning highly depends on the used document collections and the input conceptualization. We have to ensure that the document sets contain approximately the same amount of textual data (cf. Section 4.1). This guarantees that the absolute number of terms is comparable in the TF measure. When using the TF/IDF measure the absolute numbers are relativized through the size of the corpus, hence the size of both corpora must not necessarily be similar.

3.3 Human Cross-Validation

The evaluation of the results of pruning cannot only be based on measures like size and other statistical characteristics of the output. Instead, an empirical evaluation by subject specialists who assess the output has to be carried out. Only subject experts can evaluate the relevance of the extracted concepts and of their descriptiveness towards the specified domain. It is impossible to evaluate each individual output in practise.

Therefore, we base the assessment on the comparision of the pruning output with a gold-standard ontology which includes only the concepts from the source ontology that have determined to be domain relevant by the subject specialists. Thereby we can study the effects of different parameters with respect to overlap between pruned and assessed ontologies.

4 Evaluation Results

We evaluated the pruning technique within the UN FAO AOS project. The target domain of the output ontology has been biosecurity. This domain involves aspects like food safety, animal health and plant health. We have reused a general thesaurus on agricultural terms as input ontology.

4.1 Pruner Input

4.1.1 Domain Corpus

Three sets of documents have been compiled for evaluation purposes by subject specialists, which capture the above mentioned sub aspects of biosecurity. The domain corpus contains 90 documents and is 9.73 MB large.

Concepts	Relation Types	Relations Hierarchical	Relations Non-Hierarchical	Relations Related Terms	Relations Used For	Maximum Tax. Depth
17506	3	17168	15285	13486	1799	8

Table 1: AGROVOC Thesaurus Statistics

4.1.2 Generic Corpus

Two different generic document sets have been compiled:

Generic Corpus 1 (Gen): The first set of generic documents has been chosen randomly from generic news sites and the reuters 21578 test collection [7]. It contains 32 documents accounting a total of 9.55 MB of data.

Generic Corpus 2 (AG): A second generic document set has been chosen to test the behaviour of our approach when comparing the domain corpus with a corpus from a similar domain. This second generic corpus has been compiled out of randomly chosen html news articles from the US Department of Agriculture, documents from different FAO research areas, hence covering a broad range of agricultural topics. This adds up to a collection of 215 documents at a size of about 4 MB.

4.1.3 Input Ontology

We used the UN FAO AGROVOC thesaurus as input for the evaluation (cf. Table 1 for statistics). AGROVOC is a thesaurus and contains 3 relation types, which are frequently instantiated: ”related terms” expresses arbitrary relationships between concepts, ”used for” expresses that one concepts is used as a descriptor of the hyponym relationship which constitutes a taxonomy. The taxnomy of AGROVOC is a rather flat structure with respect to the high number of concepts, since the maximum depth is 8.

4.2 Evaluation Settings

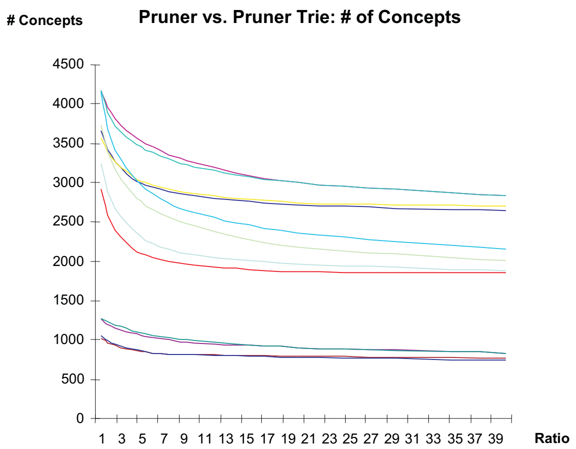
We carried out two evaluations. First, both Vector and Trie-based concept identifications have both been evaluated using corpus *Gen* with varying frequency weights and granularities. The ratio has been varied using the discrete values (1.0, 2.0, 4.0, 6.0, 10.0, 20.0, 40.0).

The pruned ontology with the highest number of concepts has been chosen for empirical assessment and evaluation by subject specialists. Subject specialists deleted all concepts in the pruned ontology that were not relevant for the domain. This evaluated ontology has been used as the gold-standard ontology. All other pruned ontologies have been compared with it testing the effects of different parameter settings on the filtering of relevant and more specific concepts.

4.3 Statistical Results

4.3.1 Trie vs. Vector-based concept identification

The peformance of both identification techniques is shown in Figure 2. Obviously, 3 clusters or groups of curves can be identified. The upper 4 curves represent the results



Ratio

	1	2	4	6	10	20	40
Trie_TFIDF_ALL_Gen	3660	3350	3063	2948	2828	2718	2644
Trie_TFIDF_ONE_Gen	4167	3892	3602	3450	3255	3005	2830
Trie_TF_ALL_Gen	3568	3319	3078	2969	2855	2751	2699
Trie_TF_ONE_Gen	4176	3802	3530	3384	3192	2996	2834
Pruner_TF_ONE_Gen	1278	1177	1091	1033	962	900	830
Pruner_TF_ALL_Gen	1019	954	872	841	818	792	762
Pruner_TFIDF_ONE_Gen	1277	1213	1120	1065	991	906	838
Pruner_TFIDF_ALL_Gen	1058	963	885	837	814	779	752
Trie_TF_ONE_AG	4145	3540	3100	2864	2615	2359	2149
Trie_TF_ALL_AG	3246	2771	2422	2227	2086	1964	1883
Trie_TFIDF_ONE_AG	3737	3281	2871	2662	2443	2177	2007
Trie_TFIDF_ALL_AG	2909	2488	2160	2050	1946	1864	1847

Trie_TFIDF_ALL_Gen
Trie_TFIDF_ONE_Gen
Trie_TF_ALL_Gen
Trie_TF_ONE_Gen
Pruner_TF_ONE_Gen
Pruner_TF_ALL_Gen
Pruner_TFIDF_ONE_Gen
Pruner_TFIDF_ALL_Gen
Trie_TF_ONE_AG
Trie_TF_ALL_AG
Trie_TFIDF_ONE_AG
Trie_TFIDF_ALL_AG

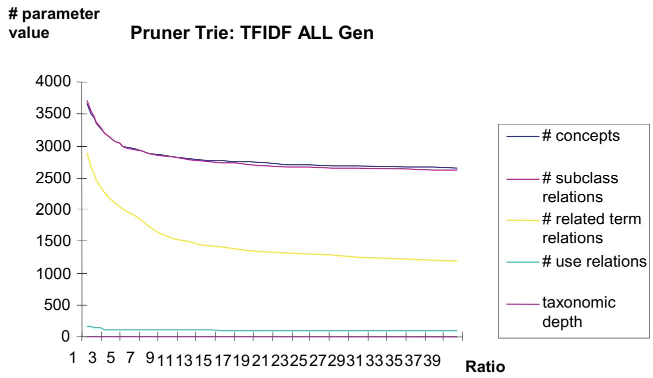
Figure 2: Vector-based vs. Trie-based concept identification

of the Trie-based concept identification and generic corpus (Gen). The curves in the middle belong to usage of generic corpus (AG), whereas the lower 4 curves show the results of vector-based concept identification.

Subset tests show, that all ontologies obtained via vector-based concept identifications are a total subset of the Trie-based identification. Obviously more concepts can be recognized when compound words can be used.

4.3.2 Influence of user parameters

Within all three groups of curves, two sub groups can be identified. Granularity ONE always identifies more concepts than granularity ALL. The usage of TFIDF vs. TF as frequency measures makes no significant difference in our evaluation. This can be



Ratio	1	2	4	6	10	20	40
# concepts	3660	3350	3063	2948	2828	2718	2644
# subclass relations	3709	3380	3064	2938	2812	2684	2605
# related term relations	2889	2455	2086	1908	1536	1329	1183
# use relations	143	134	122	119	105	97	86
taxonomic depth	7	6	6	6	6	6	6

Figure 3: Effects of varying multiplicity ratio

accounted to the fact that both weights are relativized through the comparsion of domain and generic frequencies.

The multiplicity ratio monotonically decreases the number of identified concepts. The minimal set of concepts is constituted by those concepts that can only be identified in the domain corpus and do not occur in the generic corpus at all.

Closer study of the effects of varying the multiplicity ratio (cf. Figure 3) shows that the development of the hierarchical relationships and the 'related terms' relationships almost directly correlates with the number of concepts, whereas the 'use' relationship and the taxonomic depth do not vary significantly, in fact show very little decrease only.

4.3.3 Influence of generic corpora

The use of (AG) corpus leads to smaller ontologies containing an average of 2565 concepts versus an average of 3234 concepts using the Gen corpus. Subset tests show that none of the pruned ontologies resulting from the AG set is a complete subset of its Gen counterpart. On average the AG outputs contain 213 concepts (with a standard deviation of 53) which are not found in the Gen output. On the other hand, an average of 883 (with a standard deviation of 235) concepts have been pruned using AG instead of Gen. This number is quite constantly distributed amongst all outputs. On the other hand an average of 2351 concepts could be identified using both corpora.

5 Conclusion

Our results clearly show that the ability to recognize compound words drastically improves the results. Manual inspection of the pruned ontologies also shows that generic corpora closely related to the intended target domain such as AG leads to a bigger upper-level of the ontology, i.e. allow to generalize the resulting ontology.

The evaluation has been based on the largest resulting ontology, which has been automatically extracted from the ontology, given the used parameter variations.

It would be interesting to see, if the largest pruned ontology actually contains all concepts that are identified by an exhaustive manual assessment of the input ontology itself. Given the restrictions of time and cost, however, this is unrealistic. A first empirical manual assessment [5] has shown that a generic document set, which represents the surrounding area of the target domain (here the AG set), succeeds in identifying more of the non-relevant concepts. This higher rate could hence only be achieved on a higher total cost of losing a larger set of domain relevant concepts.

In conclusion, no clear statement can be derived concerning an optimal parameter setting. If the aim is to extract possibly all relevant information from the source ontology, then the best approach is to apply the pruner with the least restrictive parameter setting and then further assess the result by subject experts. If, however, subject experts are not available and the goal is to rather retrieve a subset of the source ontology, which includes the least possible amount of irrelevant concepts, even on risk of loosing valuable concepts, then a more restrictive set of parameters should be chosen.

The experience collected with using different generic corpora, shows that a slightly different compilation of the document sets might lead to different results. For our application it might therefore be interesting to identify three different domain document sets representing each sub areas of the target application, viz. food safety, animal health and plant health, separately and apply them to the pruner in separate evaluation runs, later merging the resulting ontologies. In further work, this evaluation should be applied in different domains, in order to see if the statements and conclusions derived above still hold.

Acknowledgements

We thank the UN FAO, Rome, for substantial contributions to and the funding of our work. We also thank Andreas Hotho (AIFB, University of Karlsruhe) and Boris Motik (FZI) for their technical guidance.

References

1. J. Davies, D. Fensel, and F. van Harmelen. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. John Wiley and Sons, 2003.
2. M. Fernández-López, A. Gómez-Pérez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using Methontology and the Ontology Design Environment. *Intelligent Systems*, 14(1), January February 1999.
3. E. Fredkin. Trie memory. *Communications of the ACM*, 3(9):490–499, 1960.
4. K.S. Jomes. A statistical interpretation of terms specificity and its application retrieval. *Journal of Documentation*, 28:11–20, 1972.

5. B. Lauser. Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment. Diplomarbeit, University of Karlsruhe, Institut AIFB, D-76128 Karlsruhe, Germany, 2003.
6. B. Lauser, T.Wildemann, A. Poulos, F. Fisseha, J. Keizer, and S. Katz. A comprehensive framework for building multilingual domain ontologies: Creating a prototype biosecurity ontology. *DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence, Florence, Italy*, October 2002.
7. D.D. Lewis. Reuters-21578 text categorization test collection distribution 1.0. Internet: http://www.research.att.com/#lewis, 1999.
8. G. Salton. *Automatic Text Processing: The transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1988.
9. S. Staab and al. GETESS - searching the web exploiting german texts. In *Proc. of. Cooperative Information Agents (CIA)*, pages 113–124, 1999.
10. S. Staab, H.-P. Schnurr, R. Studer, and Y. Sure. Knowledge processes and ontologies. *IEEE Intelligent Systems, Special Issue on Knowledge Management*, 16(1), January Febrary 2001.