Optimizing ontology learning systems that use heterogeneous sources of evidence

Gerhard Wohlgenannt, Stefan Belk, and Katharina Rohrer

Vienna Univ. of Economics and Business, Welthandelsplatz 1, 1200 Wien, Austria
{gerhard.wohlgenannt,stefan.belk,karohrer}@wu.ac.at
http://www.wu.ac.at

Abstract. As the manual construction of ontologies is expensive, many systems to (semi-)automatically generate ontologies from data have been built. More recently, such systems typically integrate multiple and heterogeneous evidence sources. In this paper, we propose a method to optimize ontology learning frameworks by finding near-optimal input weights for the individual evidence sources. The optimization process applies a so-called source impact vector and the Tabu-search heuristic to improve system accuracy. An evaluation in two domains shows that optimization provides gains in accuracy of around 10%.

Keywords: heterogeneous evidence sources, ontology learning, optimization, spreading activation

1 Introduction

Ontologies are a cornerstone of the Semantic Web. As the manual construction of ontologies is expensive and cumbersome, systems for (semi-automatic) learning of ontologies have been created, which bootstrap the ontology construction process using data-driven methods. Naturally, as the task at hand is very complex, automatically generated ontologies are (i) typically lightweight (containing few axioms), and (ii) contain correct, but also wrong, constituents. Therefore an obvious goal in ontology learning is improving system accuracy. In contrast to seminal work on ontology learning, which used a single domain text corpus to extract facts, more recently there has been some work which uses multiple and heterogeneous sources (see also next section). Using multiple sources can provide accuracy gains, as it can better exploit redundancy of facts found in different sources. Redundancy of evidence in various sources can be seen as a measure of trust and relevance [5].

In previous work, we studied – in an ontology learning context – how many sources are necessary to benefit from heterogeneous sources, and how much evidence is sufficient per single source [11]. This research was an important step to find guidelines on how to configure an ontology learning system regarding the number of evidence sources, and Wohlgenannt [11] deliberately used the same input weights for all sources – in order to isolate the effect of using multiple and heterogeneous (unstructured, semi-structured, and structured) sources.

In this paper, we build on previous work, and aim to further optimize system accuracy by using an optimization algorithm (Tabu search [7]) to find the best combination of input weights for the individual evidence sources.

The research questions are as follows: (i) How well can an ontology learning system be optimized by adapting source input weights? – especially if quality of evidence varies between sources. (ii) What is the influence of the number of sources used in the system on the optimization results? (iii) What other findings and guidelines can be extracted from the data collected in the optimization runs?

To address the research questions, we did two batches of optimization runs. The first one was conducted in 2013 with all 32 evidence sources used in our system, which was not very well tuned at that point. The second set of optimization runs was done in 2015, then with a better tuned system, and a reduced set of evidence sources (according to our findings in our previous work [11] that a limited number of sources is sufficient for high accuracy).

The structure of this paper is as follows: Section 2 discusses related work, and Section 3 provides an overview of the ontology learning system used, and of the heterogeneous evidence sources. Then we introduce the optimization method (Section 4), and evaluate the optimization process with different configuration settings in Section 5. Section 6 concludes the paper.

2 Related Work

There has been a lot of initial effort in building ontology learning systems around the year 2000. These systems usually process high quality domain text with statistics- and linguistics-based methods. For example, Text2Onto [3] combines machine learning approaches with linguistic processing to build so called Probabilistic Ontology Models. These models are independent of a concrete target language and attach probabilities to learned ontological structures.

Obviously, there are also more recent efforts to build systems to learn ontologies from text, for example OntoGain [6], which uses a number of algorithms such as formal concept analyses or association rule mining for unsupervised ontology construction. Other systems often try to compute a probability for ontological elements learned, which is used in the selection and integration process. For example, Abeyruwan et al. [1] suggest a method for unsupervised bottom-up ontology generation which selects ontological elements by a respective Bayesian probability. For details and information on other ontology learning systems see the recent survey publication by Wong et al. [14].

There have been few efforts yet to learn ontologies from heterogeneous evidence sources. Manzano-Macho et al. [5] outline some of the potential benefits of using heterogeneous evidence sources: the combination of sources leads to an overlap of the ontological elements suggested, this redundancy can be seen as a measure of relevance and trustiness for a certain domain. And, obviously, some methods will add valuable complementary information that other sources or methods did not detect. In their approach to combine heterogeneous sources

of evidence, Manzano-Macho et al. [5] aim at building a taxonomy of concepts with higher accuracy compared to using a single evidence source only.

Another method for the integration of heterogeneous sources has been proposed by Cimiano and Völker [2]. They focus on learning taxonomic relations between concepts combining multiple methods, and then convert the evidence found into first order logic features. Standard classifiers are applied to find useful combinations of evidence sources.

In this paper there is no focus on improving single evidence sources, but rather on the smart combination of heterogeneous evidence sources. Existing systems typically use a domain corpus as input, whereas our framework integrates a wider variety of sources, including evidence from APIs available on the Social Web and a Linked Data source. The sources are heterogeneous regarding a number of aspects: quality, type, number of evidences, etc. (see also Section 3.1).

We use spreading activation as our main method to integrate evidence. Spreading activation is a method to search neural and semantic networks, its suitability for information retrieval tasks has been demonstrated eg. in [4].

3 The Ontology Learning Framework

All experiments presented in the paper are based on an existing ontology learning system which evolved over the years. This section can only give a quick overview of the framework, for details on the workings of the system see Wohlgenannt et al. [12], or Liu et al. [9], who present the original version of the framework. In a nutshell, the system starts from a (typically small) seed ontology, which is extended with additional concepts and relations. So the main tasks are the selection of new concept candidates and the positioning of concepts with regards to the seed ontology.

We compute new ontologies for the domains in question in regular intervals (monthly) to trace the evolution of the domains. Figure 1 shows (parts of) the graphical representation of an example ontology learned in the *climate change* domain on data from January 2014. The concept colors indicate the ontology extension stage, the darker, the earlier the concepts were introduced. Before explaining the system in more detail, we take a look at the evidence sources used in the process.

3.1 The Evidence Sources

As the name suggests, the *evidence sources* provide the data needed to extend (learn) ontologies. In general, the input to evidence acquisition is a term (typically the label of a seed concept), and the result is a list of terms related to the seed term, and optionally significance values. So, for example, the system sends the seed concept label "CO2" to the Flickr API[1], and gets a list of related terms – which will then be used in the ontology learning process together with information from all other sources of evidence.

[1] www.flickr.com/services/api/flickr.tags.getRelated.html

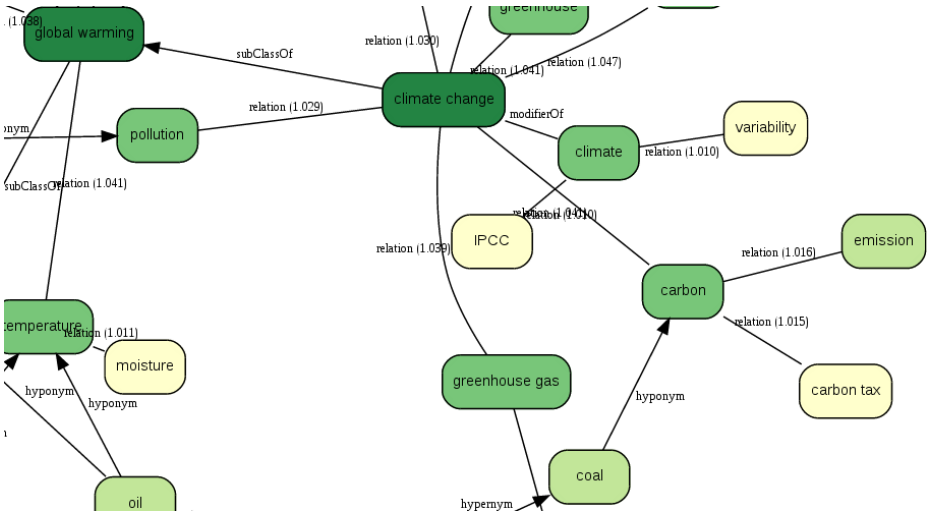

Fig. 1. Part of a sample ontology in the domain of climate change after three stages (levels) of extension

The evidence sources are heterogeneous regarding various aspects, such as (i) the number of evidences returned; (ii) the average quality (ie. domain relevance) of terms provided; (iii) the update frequency of the source – some sources are dynamic, eg. social sources and news media, some rather static, eg. WordNet and DBpedia; (iv) the availability of a significance score for the terms; and (v) the type of underlying data (text, structured, etc.).

By default, the ontology learning system currently uses 32 sources of evidence, which are listed in Tables 1 and 2. The first batch of experiments (see Section 5) conducted in year 2014 is based on all 32 sources, the second batch (year 2015) is based on the sources marked with boldface fonts. A major fraction of evidence sources is made up by the 16 sources based on keywords computed with co-occurrence statics on documents published and mirrored in the respective period of time, and filtered with a domain-detection service. The system computes page- and sentence-level keywords for documents collected from: US news media, UK news media, AU/NZ News Media, Websites of NGOs, Fortune 1000 company Websites, Twitter tweets, Youtube postings, Google+ postings, and public Facebook pages and postings. Furthermore, we use Hearst patterns [8] on those corpora, which constitutes further 10 evidence sources. Currently, the system includes two evidence sources based on calls to APIs of Social Media sites (Twitter, Flickr). Structured evidence sources contribute the remaining 4 sources of evidence, ie. hypernyms, hyponyms and synonyms from WordNet, and related terms from DBpedia. For more details on evidence sources used see Wohlgenannt et al. [13].

	Extraction Method		
Data sources			
domain text from:	Keywords/page	Keywords/sentence	Hearst patterns
US news media	**1**	2	**3**
UK news media	4	**5**	6
AU/NZ news media	7	**8**	9
Other news media	**10**	11	12
Social media – Twitter	13	-	**14**
Social media – Youtube	**15**	-	16
Social media – Facebook	17	-	18
Social media – Google+	19	-	20
NGOs Websites	21	22	**23**
Fortune 1000 Websites	**24**	25	26

Table 1. The first 26 evidence sources are based on domain-specific text collected from the Web.

	Method				
Data source:	hypernyms	hyponyms	synonyms	API	SPARQL
WordNet	**27**	**28**	29	-	-
DBpedia	-	-	-	-	**30**
Twitter	-	-	-	**31**	-
Flickr	-	-	-	**32**	-

Table 2. The other 6 sources of evidence, based on WordNet, DBpedia and Social Media APIs.

3.2 Source Impact

As mentioned, evidence sources are heterogeneous in number and quality of terms provided, we use a so-called Source Impact Vector (SIV) to manage the influence of a particular evidence source on the ontology learning process. Equation 1 demonstrates that the SIV consists of one impact value per evidence source (and point in time). The impact value is in the interval $[0.0, 1.0]$, a value of 1.0 results in high impact in the learning processes, wheres 0.0 in fact omits evidence suggested by the respective source.

$$SIV_t = \left[I_{es_1,t} \;\; I_{es_2,t} \;\; \cdots \;\; I_{es_n,t} \right] \tag{1}$$

The SIV is used to set the weights in the spreading activation network (see next subsection for details), which selects new concept candidates for the ontology. Initial versions of the system ([9], [10]) applied a manually picked and static source impact, in this paper we propose novel ideas and experiments to optimize the ontology learning system via the SIV. The optimization process aims to find a configuration of the SIV which maximises the ratio of relevant new concept candidates suggested by the system.

3.3 The Ontology Learning Process

Having described the evidence sources and the SIV, we can introduce the basic workflow of the ontology learning system: The ontology learning run starts with a small seed ontology (in the climate change usecase we use two concepts, namely *climate change* and *global warming*). The system collects evidence for the seed concepts from the 32 (or 14) evidence sources. After integrating all this evidence into a semantic network, a transformation process using the SIV converts the semantic network into a spreading activation network. The spreading activation algorithm yields new concept candidates. We currently pick the 25 candidates with the highest level of activation. The concept candidates are evaluated for domain relevance by domain experts, this is the only part of the system that involves human intervention. Finally, the system positions new concepts rated as relevant with regards to the existing ontology. For the next ontology extension step the framework uses the result from the previous iteration as new seed ontology, and further extends it. We typically do three ontology extension iterations. Figure 2 gives on overview of the workflow.

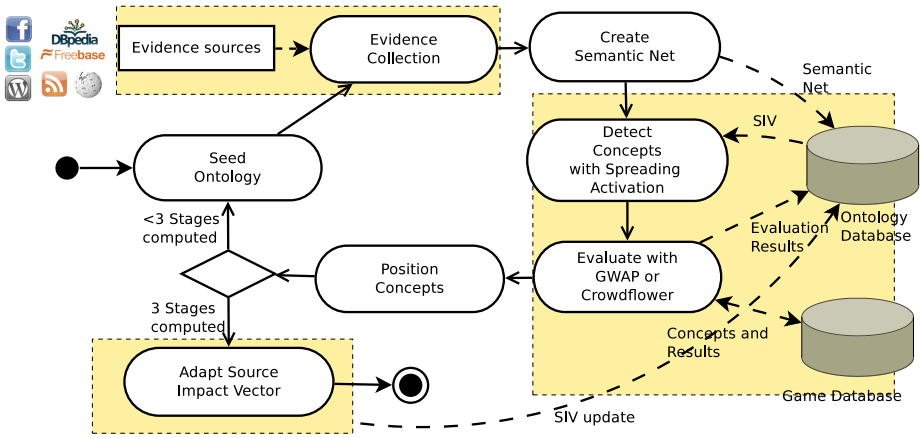

Fig. 2. The Ontology Learning Framework

The goal of this research (and the optimization) is to improve the ratio of relevant to non-relevant concept candidates, ie. to improve the output of the spreading activation algorithm. The SIV is a key factor in this optimization process, as it determines – in combination with significance scores provided by the evidence sources – the weights in the spreading activation network.

4 The Optimization Process

Although a spreading activation network has the fundamental characteristics of a neural network, we did not find a way to apply classic neural network learning techniques to optimize the output for a number of reasons:

- The spreading activation network doesn't have an explicit output layer, the *results* of the spreading activation algorithm are the activation levels of nodes all over the network.
- We select a preset number of nodes (eg. 25) with the highest activation level as concept candidates. The use of an error function (as used eg. in backpropagation) is not straightforward, as we only assess the preset number of nodes with the highest activation, but any other node might be a relevant domain concept as well. So there is no distinct correct output of the spreading activation network that could be used.
- The learning algorithm can not freely optimize the weights in the network, as values of the SIV are only factors in the connection weights. First of all, when multiple SIV factors make up a connection weight, it is not clear which specific SIV factor should be changed. And more importantly, if a specific SIV value is changed for one connection, it needs to be changed simultaneously everywhere in the spreading activation network wherever used, leading to unpredictable effects.

The characteristics described above led us to experiment with heuristics to improve the output of the ontology learning framework based on the modification of the SIV. This includes a baseline with a static SIV (Section 4.1), and a model that aims to optimize the SIV (Section 4.2).

Overall, the crucial factor which has an impact on the results of the ontology learning process are not so much the absolute values in the SIV, but the differences between evidence sources. Higher source impact for an evidence source results in increased activation levels and therefore a higher chance of being a candidate concept for evidence suggested by the particular source.

4.1 Static Source Impact Values

The simplest way to use the SIV is to have static values for any source, not changing over time or across domains. We use a source impact of 0.2 for all 32 evidence sources. This uniform source impact has been used in the experiments regarding the number and balancing of evidence sources presented by Wohlgenannt [11], and provides good results, which we use as a baseline and starting point of the optimization experiments.

4.2 Optimization

With this strategy, instead of having a single static SIV, the system investigates different SIV settings and their results. In the first batch of experiments, we set

the source impact for every evidence source to values in the interval $[0.0, 1.0]$ with a step-size of 0.1, i.e. *eleven* values per source. With 32 evidence sources, this leads to an enormous number (11^{32}) of potential permutations. As a single ontology learning run (depending on settings) takes around four hours of computation time, we decided to use the Tabu Search heuristic [7], and simply optimize every evidence source by itself, with settings for other sources constant. The leads to 352 ($11 * 32$) ontology learning runs. In the second batch, we used a step-size of 0.2 and 14 evidence sources, resulting in 84 ($6 * 14$) ontology learning runs.

The following Pseudocode shows the Tabu Search-based optimization strategy:

Algorithm 1 Optimize SIV with Tabu Search

```
Initial solution ← Static SIV
# do for all 32 evidence sources
for each evidence source e do
  # create neighborhood
  for X in interval [0.0, 1.0] size 0.1 do
    # evaluate every neighbor
    SIV_e ← X
    compute ontology (all 3 extension steps) using SIV
    Q_x ← Evaluate quality of ontology
    Remember result (X, Q_x)
  end for
  # keep value X with best result – skip the rest
  SIV_e ← pick best result from neighborhood
  Put all other solutions from neighborhood on Tabu list
end for
```

Basically, the heuristic looks for the *best* source impact value for a single evidence source, and uses this value when optimizing the other evidence sources. One of the downsides of this method is that the order in which evidence sources are processed obviously has an impact on the result. The system randomizes the order of evidence sources before every optimization run. It will typically not find a global optimum, but hopefully a good solution with a limited number of permutations. Furthermore, the optimization helps to visualize and understand how specific SIV settings contribute to ontology quality.

5 Evaluation

This section summarizes the findings of optimization runs performed to gain insights about the improvements of accuracy which can be reached by optimizing the combination (the impact) of evidence sources.

5.1 Evaluation Setup

In previous experiments conducted in year 2014 (see Section 5.3 for results) we used a step-size of 0.1 in a source impact interval of 0.0 to 1.0, and 32 evidence sources. For the recent batch of evaluation experiments we used a more computationally efficient setup, with optimization runs for 14 evidence sources, and a step-size of 0.2. Previous work shows that 10-15 evidence sources are sufficient to have good results [11].

Relevance assessment of concept candidates is being done by domain experts. The accuracy values used in this section are simply the number of concept candidates rated as domain-relevant by the domain experts divided by all concept candidates suggested by the system. We decided to use the ratio of relevant concepts as evaluation metric because (i) the relevance of domain concepts is critical to generating useful domain ontologies, and (ii) relevance assessment for concept candidates is the only part of the system where manual input is applied.

5.2 Recent Optimization Experiments

These experiments were conducted with the latest version of the ontology learning system in the first half of year 2015. We compared the results of using a static SIV (uniform source impact of 0.2 for all evidence sources) to optimizing source impact.

Domain	Static SIV	Optimized	Improvement
Climate Change (en)	67.15%	76.88%	9.73%
Tennis (en)	44.57	54.42%	9.85%

Table 3. System accuracy and gains by optimizing the SIV as compared to a static SIV, in two different domains.

Table 3 summarizes the results for two different domains, the domains of *climate change* and of the sport *tennis*. The values in the table represent the average accuracy for ontology generation runs with a static SIV and for the optimization processes. The data indicates a substantial improvement in accuracy of around 10% which can be reached by optimizing the SIV.

The difference in accuracy between the two domains can be attributed to the following reasons: (i) In the *climate change* domain we are supplied with much bigger and domain-specific corpora, whereas with *tennis* we use general news corpora which are then filtered for tennis-related documents ex-post. Besides corpus-size and quality, (ii) the domain of tennis has a lot of overlap with other sports domains. Concepts such as ball, tournament, etc. attract related but not domain-relevant terms from other sports, whereas climate change seems to be more "closed".

However, the most interesting fact is the improvement in accuracy, which is statistically significant, as confirmed with a binomial test.

5.3 Previous Experiments

This section discusses experiments done in early 2014 with an older version of the system, which wasn't as well tuned, with lower general accuracy. This is reflected by the accuracy values for *Static SIV* in Table 4. We evaluated two settings, where we used either up to 50 (*limit=50*) evidences per source and concept, or up to 200 (*limit=200*)[2] – for details on these settings see Wohlgenannt [11].

Domain	Static SIV	Optimized	Improvement
Climate Change (en) – *limit=50*	63.33%	78.18%	14.85%
Climate Change (en) – *limit=200*	64.13%	77.33%	13.20%

Table 4. System accuracy of the previous system version, in the domain of *climate change*, for two settings.

In year 2014 we started from a lower baseline (around 63-64%), and experience improvements from optimization between 13-15%, more than in the recent batch of experiments. Our interpretation of the results is the following:

- The lower baseline leaves more room for improvement.
- In the 2014 experiments we used a *step-size* of 0.1, which resulted in higher computational cost, but also a more fine-tuned optimization.
- The number of evidence sources was much higher (32 sources), therefore the potential for fine-grained optimization of sources was higher.

With regards to the research questions posed in Section 1, the evaluation shows that system accuracy can be raised substantially by optimization using the SIV. It helps to have a high number of evidence sources and also a fine-grained step-size, this allows for a more precise optimization process.

5.4 Analysis of Evidence Sources

The evidence sources provide terms and relations of different quality to the learning algorithms. Wohlgenannt [11] discusses the quality and characteristics of evidence sources in some detail. In a nutshell, the number and quality (domain relevance) of evidence is very heterogeneous. Keyword-based sources typically provide a high number of terms, with good quality for the terms with highest co-occurrence significance, but degrading with more terms added having a lower significance. Terms for structured sources such as DBpedia and WordNet generally offer good quality, but low term numbers. In our experiments, APIs of social sources such as Twitter and Flickr yield mostly low quality terms – but we still have them included to (i) benefit from the effect of redundancy between sources, and (ii) as they often provide very recent and complementary terminology.

[2] The more recent evaluations in Section 5.2 were conducted with *limit=50* settings.

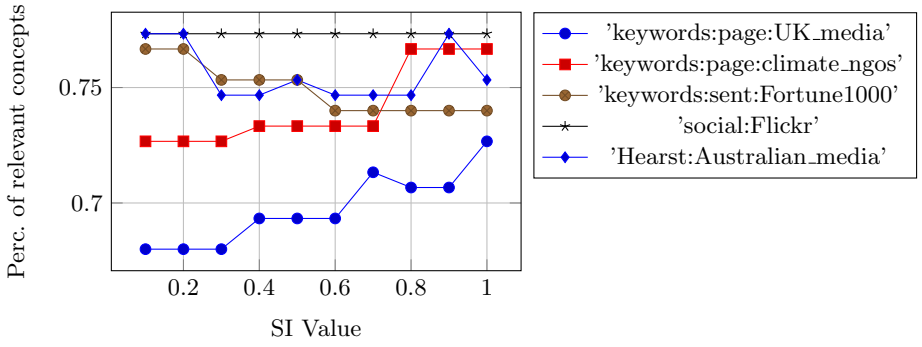

Fig. 3. Influence of Source Impact settings for a number of selected evidence sources.

Figure 3 visualizes the influence of source impact (SI) settings for some individual sources on system accuracy. The data is taken from an optimization run in the domain of *climate change*, and helps to explain the characteristics and experiences with SIV optimization. Usually evidence sources fall into one of the following categories:

- *Increasing the SI raises accuracy.* These evidence sources obviously yield relevant terms and helpful contributions to the ontology learning system. With higher impact of the source the accuracy goes up. `keywords:page:UK_media` and `keywords:page:climate_ngos` in Figure 3 fall into this category.
- *Increasing the SI lowers accuracy.* This applies to sources which do not contribute much helpful data. For example `keywords:sent:Fortune1000`.
- *Accuracy independent of SI.* This usually happens when a source provides a very low number of evidences, `social:Flickr` in our example.
- *Erratic.* As with `Hearst:Australian_media`, sometimes the effects of the SI are rather erratic. Such cases are the biggest challenge for optimization.
- A mix of the basic categories described above.

Erratic behavior or a mix of the categories described above results from the fact that the system selects the 25 concept candidates with the highest spreading activation level. Raising the influence of a single evidence source gives more importance to all its evidence, relevant or not. The Tabu search heuristic will not find an optimal, but typically good, combination of sources (ie. the SIV).

6 Conclusions

Ontology learning aims at (semi-)automatically constructing lightweight ontologies from sources of evidence. When using multiple and heterogeneous sources, balancing and optimizing the influence of evidence sources is crucial. In this paper, we introduce and evaluate a strategy for optimizing such ontology learning systems, and see improvements in accuracy (in the concept detection phase) of

ca. 10-15%. The contributions are as follows: (i) Presenting a novel method to configure and optimize ontology learning systems using the source impact vector and the Tabu-search heuristic, and (ii) experiments in two domains to estimate the accuracy gains from this optimization technique. Future work includes the repetition of experiments in other domains, also based on corpora in other languages, and the application of alternative optimization strategies.

Acknowledgments. The work presented was developed within uComp, a project which receives the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET.

References

1. Abeyruwan, S., Visser, U., Lemmon, V.P., Schürer, S.C.: Prontolearn: Unsupervised lexico-semantic ontology generation using probabilistic methods. In: Fernando Bobillo, e.a. (ed.) URSW. LNCS, vol. 7123, pp. 217–236. Springer (2013)
2. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Ontology Learning from Text, chap. Learning Taxonomic Relations from Heterogeneous Sources of Evidence, pp. 59–76. IOS Press, Amsterdam (2005)
3. Cimiano, P., Völker, J.: Text2onto. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB. Lecture Notes in Computer Science, vol. 3513, pp. 227–238. Springer (2005)
4. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)
5. David Manzano-Macho, A.G.P., Borrajo, D.: Unsupervised and domain independent ontology learning: Combining heterogeneous sources of evidence. In: Nicoletta Calzolari, e.a. (ed.) LREC'08. ELRA, Marrakech, Morocco (May 2008)
6. Drymonas, E., Zervanou, K., Petrakis, E.: Unsupervised ontology acquisition from plain texts: The ontogain system. In: Hopfe, C. (ed.) Natural Language Processing and Information Systems, LNCS, vol. 6177, pp. 277–287. Springer (2010)
7. Glover, F., Laguna, M.: Tabu Search. Kluwer, Norwell, MA, USA (1997)
8. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING'92. pp. 539–545. Nantes, France (1992)
9. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management 0(1), 50–58 (2005)
10. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. Data & Knowledge Engineering 69(8), 763–778 (2010)
11. Wohlgenannt, G.: Leveraging and balancing heterogeneous sources of evidence in ontology learning. In: Fabien Gandon, e.a. (ed.) ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. LNCS, vol. 9088, pp. 54–68. Springer (2015)
12. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Confidence management for learning ontologies from dynamic web sources. In: Proceedings of KEOD 2012. pp. 172–177. SciTePress, Barcelona, Spain (October 2012)
13. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Dynamic integration of multiple evidence sources for ontology learning. Journal of Information and Data Management (JIDM) 3(3), 243–254 (2012)
14. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM Computing Surveys 44(4), 20:1–20:36 (Sep 2012)