DC Proposal: Ontology Learning from Noisy Linked Data

Man Zhu[⋆,⋆⋆]

School of Computer Science & Engineering, Southeast University, Nanjing, China
mzhu@seu.edu.cn

Abstract. Ontology learning - loosely, the process of knowledge extraction from diverse data sources - provides (semi-) automatic support for ontology construction. As the 'Web of Linked Data' vision of the Semantic Web is coming true, the 'explosion' of Linked Data provides more than sufficient data for ontology learning algorithms in terms of quantity. However, with respect to quality, notable issue of *noises* (e.g., partial or erroneous data) arises from Linked Data construction. Our doctoral researches will make theoretical and engineering contribution to ontology learning approaches for noisy Linked Data. More exactly, we will use the approach of Statistical Relational Learning (SRL) to develop learning algorithms for the underlying tasks. In particular, we will learn OWL axioms inductively from Linked Data under probabilistic setting, and analyze the noises in the Linked Data on the basis of the learned axioms. Finally, we will make the evaluation on proposed approaches with various experiments.

1 Motivation

Ontology learning refers to the task of providing (semi-) automatic support for ontology construction [3], and can overcome the knowledge acquisition bottleneck brought by the tedious and cumbersome task of manual ontology construction [17]. Recent ontology learning approaches have attempted to learn ontology from various types of data sets, such as text, xml, and database, but they seldom explore learning from Linked Data. Based on URIs, HTTP and RDF, the Linked Data project [2] aims to expose, share and connect related data from diverse sources on the Semantic Web. Linked Open Data (LOD) is a community effort to apply the Linked Data principles to data published under open licenses. With this effort, a large number of LOD data sets have been gathered in the LOD cloud, such as DBpedia, Freebase and FOAF profiles. LOD has gained rapid progress and is still growing constantly. Until May 2009, there are 4.7 billion RDF triples and around 142 million RDF links [2]. After that, the total has been increased to 16 billion triples in March 2010 and another 14 billion triples have been published by the AIFB according to [21].

⋆ Advisor: Zhiqiang Gao, School of Computer Science & Engineering, Southeast University, Nanjing, China, zqgao@seu.edu.cn

⋆⋆ Advisor: Zhisheng Huang, Department of Mathematics & Computer Science, Vrije University, Amsterdam, The Netherlands, huang@cs.vu.nl

L. Aroyo et al. (Eds.): ISWC 2011, Part II, LNCS 7032, pp. 373–380, 2011.

The advantages of learning from Linked Data, and what distinguishes it from learning from other resources, are depicted in Table 1. The most common used learning resource is HTML documents, which emerged and developed ever since the invention of the Web. The distinguishing feature thereof is that they constitute a large-scale data set and are generally publicly accessible. However, the structures inside are formed through simple HTML tags, and the HTML documents are linked to each other on document level. Compared to HTML documents, XML documents (made to be the origin of comparison in Table 1) are far more easily accessible to machines. XML can overcome the shortcomings of HTML (highly human interpretable contents, not for machines) to some extent, because XML documents contain certain structural information. The characters of glossaries are similar to that of XML documents. Besides the links among words (phrases), the structures inside are simple. The biggest problem of learning from database is that it is limited in both contents and accessibility. Learners can only learn from databases of specific domain. According to the description and statistics described in the last paragraph, we conclude that compared with other resources Linked Data is superior in that it is publicly available, highly structured, relational, and large with respect to learning.

The other side of the Linked Data coin poses the challenges we are going to cope with during the doctoral research: First, due to the publishing mechanism of the Linked Data, it contains noises inherently [4,1]. Hogan et al. analyzed the types of noises which exist in the Linked Data [9]. We are particularly interested in handling two types of noises: *partiality* and *error*. Partiality means that concept assertions or the relationships between named individuals are actually true but missed, and error means that the RDF triples are not correct (with respect to some constraints). Take a family ontology for example. The declarations of 'Heinz is a father' and 'Heinz is a male' exist in the RDF triples, then Heinz should have a child, however it is not declared in the ontology. This is an example of partiality. Besides, if we know Anna has a child, and she is a female, then Anna should not be a father, but in the ontology Anna is incorrectly declared to be a father. This illustrates the error case. Second, the ontologies in Linked Data are generally inexpressive. For example, one of the most popular ontologies, DBpedia ontology[1], is claimed as a shallow ontology. The TBox of this ontology mainly includes a class hierarchy [10].

In our doctoral researches, we endeavor to inductively learn ontologies using statistical relational learning (SRL) models. The development of SRL has been driven by real-world needs of handling noises, relations (Figure 1). In the early days, ML community have been focused on learning deterministic logical concepts. However, those methods failed to fit perfectly for noises and large-scale circumstances, which leads to statistical methods that ignored relational aspects of the data, such as neural networks, generalized linear models. On the other hand, inductive logic programming (ILP) is designed to learn first-order rules directly which are much more expressive [18]. It is argued in [4] that inductive learning methods, could be fruitfully exploited for learning from Linked Data.

[1] http://wiki.dbpedia.org/Ontology

Table 1. Comparison of learning from various resources. XML Document is made to be the origin of comparison, indicated by '0'. '+' and '−' denote degree of the corresponding character (above/below the origin), '++' and '−−' denote stronger degree than '+' and '−'.

	Publicly Available	Structured	Linked	Large
HTML Document	+	−−	0	++
Glossary	0	−	0	0
XML Document	0	0	0	0
Database	−	+	+	−
LOD	+	+	+	+

For the last few years, the ILP community and the statistical machine learning community have been incorporating aspects of the complementary technology (machine learning, probability theory, and logic), which leads to the emerging area of SRL. It attempts to represent, reason, and learn in domains with complex relational and rich probabilistic structure [8]. Using SRL, two characters of Linked Data, which distinguish Linked Data from other data sets, can be easily handled: 1) Linked Data are highly structured due to the relations between entities and the underlying ontology. 2) Linked Data contains noises, here, as described above, we refer particularly to partiality and error.

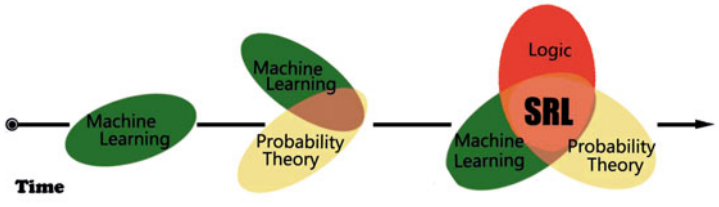

Fig. 1. The evolution of Statistical Relational Learning (SRL). SRL integrates technologies from machine learning, probability theory, and logic.

2 Related Work

There is an important body of previous work that our work builds on. We note a subset of them here. Lehmann J. et al. have done a series of work on learning Description Logics, and the algorithms proposed are implemented in DL-Learner. DL-Learner is a framework for learning Description Logics and OWL from positive (and negative) examples (in ILP, ground literals of target concept are called examples, if the ground literal is true, it is positive, negative on the contrary), and supports several learning algorithms (CELOE, random guesser learning algorithm, ISLE, brute force learning algorithm) based on ILP and machine learning [11]. [13] learned Description Logics $\mathcal{ALC}$, and [14] learned Description Logic $\mathcal{ALCQ}$ using a learning algorithm based on refinement operators, and the algorithm is implemented and evaluated in the DL-Learner. AutoSPARQL is a most

recent work, which makes use of the individual assertions in the ABox, and can be used to learn descriptions for individuals [12]. [19] proposes a log-linear Descriptions Logics based on $\mathcal{EL}$++. It integrates log-linear model with Description Logic $\mathcal{EL}$++, and can be used to learn coherent ontologies. In [20], Völker J. and Niepert M. propose a statistical approach, to be specific, association rule mining, for learning OWL 2 $\mathcal{EL}$ from Linked Data. Fanizzi N. et al. proposed a specific concept learning algorithm by extending FOIL algorithm, called DL-FOIL [7]. [6] works on the induction and revision of terminologies from metadata. Lisi F.A. et al. have done a series of work on learning rules. In [15] hypotheses are represented as $\mathcal{AL}$-log rules, and the coverage relations are defined on the basis of query answering in $\mathcal{AL}$-log. Correspondingly, [16] learns $\mathcal{DL}$-log rules, besides the differences in the expressive power of the target language, it also reformulate the coverage relation and the generality relation as satisfiability problems.

3 Proposed Approaches

In the doctoral researches, the following issues will be studied:

How to Learn from Noisy Linked Data? We propose to learn ontologies from Linked Data by SRL methods. Generally speaking, SRL models combine relational representations and probabilistic learning mechanisms such as graphical models. The majority of proposed SRL models can be categorized according to several dimensions: 1) the representation (logic or frame-based) formalisms. 2) probabilistic semantics (Bayesian networks, Markov networks, stochastic grammar etc.). We will learn ontologies under probabilistic setting, where the learning problem is transformed into finding an optimum axiom A satisfying certain function, such as $A = \arg\max_A P(A, E)$, E denotes all assertions (facts) in the original ontology. In SRL models, different probabilistic semantics are used for modeling the probability distributions, and random variables corresponding to assertions (containing partiality and errors) can encode probabilistic information, thus is suitable for handling noises. For example, in Markov logic, the facts and the terminology axioms correspond to a Markov network. The joint probability distribution is defined according to this Markov network, where nodes represent assertions with probabilities. In the current work (c.f. Sect. 4.2), the axioms are attached with weights. By maximizing the joint probability, the weights can be learned. Each step the candidate axiom with the largest weight is selected to be further expanded in the next step if the joint probability still increases. The process of finding the optimum axiom can be viewed as searching in a predetermined hypotheses space. This process can go in a top-down or bottom-up manner. Top-down algorithms start from the most general hypothesis $target \sqsubseteq \top$, and iterate to select one from candidates according to a performance criteria and add to the hypothesis until the stop criteria is reached. In bottom-up approaches the iteration begins at the most specific hypothesis whose right-hand is the intersection of all possible literals.

We will propose a SRL model suitable for learning from noisy data. Currently, a number of SRL models have been proposed from various research fields with

different application background. For example, Markov logic [5] is one of the most recent SRL model, which combines first-order logic with undirected graphical models (Markov networks). According to one of our recent works (c.f. Section 4.2), Markov logic can be applied to ontology learning from noisy data. Still the results can be better. We argue that the performance of applying the currently proposed SRL models to learn from Linked Data can be further improved by proposing a SRL model particularly for this task concerning the following aspects: 1) As we all know, OWL builds on Description Logic basis. Today's SRL models use either frame-based, which contains simple relations, or logic representation (e.g. FOL, which is more complex). In terms of expressing power, they are not the best fit for OWL. 2) Currently SRL models are still weak in analyzing the independencies inside the probabilistic model. However, independencies play an important role in saving parameter space as well as improving the computing efficiency, which should be studied carefully. Thus we will propose a more suitable model for learning from noisy Linked Data with the goal of improving both the learning accuracy and the learning efficiency.

How to Guide the Search? We will propose methods to structure the hypotheses space. In ILP algorithms such as FOIL [18], which learns first-order rules, the rules are generated through adding literals to the current rule. New literal can be of the form $Q(v_1, \ldots, v_r)$ (at least one v_i already exists in the rule), $equal(v_j, v_k)$, or the negation of either of the first two forms. Using this kind of approaches, it is still unknown that whether the following can be guaranteed: 1) will adding a new literal lead to a more specific concept? 2) can all concepts be traversed? Another approach, named refinement operator, defines a mapping $S \mapsto 2^S$ on a quasi-ordered space S, thus it structures the hypotheses space according to quasi-ordering relations, such as subsumption. A number of refinement operators have been proposed, such as $\mathcal{ALC}$ refinement operator [13] and $\mathcal{ALCQ}$ refinement operator [14]. The properties of refinement operators, such as (weakly) complete, ensure that if an axiom should be correct according to the Linked Data, it can be reached by the refinement operator. Current refinement operators will be improved by, firstly, at each step, the hypotheses generated by the refinement operator should be finite, and secondly, the refinement operator should be designed for OWL and its profiles according to specific models. For example, if Markov logic is chosen as the model, then in each step in the iteration, the dependencies between the candidate hypotheses should be minimized, so as to guarantee that the weights learned truly reflect the confidence of the hypothesis (c.f. Section 4.2).

How Many Partiality and Errors Are There? In [9], Hogan discusses common errors that can be systematically detected in RDF publishing. The results provide a significant basis for our motivations. We are still interested in analyzing the Linked Data more semantically. The axioms we learned contain two parts. One part of them already exist in the ontology, which can be evaluated automatically by comparing with the original ontology. The other part of them

are not in the ontology. Given the observation that the ontologies in Linked Data are generally inexpressive (c.f. Sect. 1), this part of axioms are not necessarily incorrect. We will evaluate them manually. Finally, we will have a set of correct axioms. We want to answer the question of "How many partiality and errors are there in Linked Data?". By querying the original ontology and comparing the results with the learned axioms, we will propose algorithms to analyze the data in ABox to know whether some of them are missed or some of them are wrongly stored in the Linked Data.

4 Results and Evaluation

4.1 DLP Learning from Uncertain Data

The origin of this work can be found in [22], where we focused on learning description logic programs (DLP) from explicitly represented uncertain data. DLP is an intermediate knowledge representation that layers rules on top of ontologies. DLP is an expressive and tractable subset of OWL, and plays an important role in the development of the Semantic Web. We modified the performance evaluation criteria based on pseudo-log-likelihood in the designed ILP like algorithm PIDLP. With the new performance evaluation criteria, uncertainties are handled and meanwhile DLPs can be learned. We also tested the algorithm in two datasets, and the results demonstrated that the approach is able to automatically learn a rule-set from uncertain data with reasonable accuracy. However, in many cases, uncertainties exist implicitly, such as in Linked Data. In what follows, we transfer our attention to learning from noisy Linked Data without handling explicitly specified uncertainties.

4.2 Learning $\mathcal{ALCI}$ from Noisy Data by Markov Logic

$\mathcal{ALCI}$ contains inverse role constructor in addition to the basic Description Logic $\mathcal{ALC}$. In our most recent work, we examine learning $\mathcal{ALCI}$ from noisy Linked Data, attempting to take the first step towards our first proposed approach. The procedure of learning can be viewed as searching for the optimize hypotheses (axiom) in the hypotheses space composed of all possible axioms according to certain criterion. In this work, Markov logic is used for handling noises. More specifically, hypotheses are accompanied with a weight which indicate the degree of consistency between the hypotheses and the RDF triples in the data set. In each iteration, the weights are learned with the target of joint probability maximization, and we choose the hypothesis with the largest weight. The iteration runs until performance stops to improve. We evaluate the approach on 4 data sets in addition to a small data set illustrating the functionality of learning definitions. The results demonstrate that the method performs well under noises, and is capable of learning $\mathcal{ALCI}$ with an average precision of 0.68 and recall 0.59.

4.3 Evaluation

The evaluation goes in two-fold (semi-)automatically. Firstly, using ontologies as gold standard, such as EKAW ontology, we separately treat the TBox and ABox in the Linked Data as testing and training set. We learn ontologies from the ABox, and evaluate the results learnt according to the TBox. Widely used IR measures *precision*, *recall*, and *F1-score* are adopted here. This way, the performance of the proposed approaches can be observed. Nevertheless the axioms in TBox may still not be complete [10], and the learned axioms not in the TBox are not bound to be wrong. Thus secondly, the part of learned axioms not in the TBox will be manually evaluated. We assign 3 numbers to indicate the correctness of the axioms: 1-low, 2-medium, and 3-high. For each axiom, a group of people will determine whether or not they think it is correct, and offer a judgement represented as a number. The final result will be an average.

5 Future Works

In the future, we plan to do the following works: firstly, we will further improve our approaches of learning OWL axioms from Linked Data (such as DBPedia) by adopting new SRL models. By proposing mechanisms for representing OWL axioms with probabilistic graphical models, and analyzing the independencies inside, more accurate results and more efficient algorithms can be found. Secondly, we will work towards answering the question "How many partiality and errors are there?". According to the preliminary results we got from the work (c.f. Sect. 4.2), a number of axioms we learned that are not in the Linked Data should be correct. In addition to the analyzes carried out by Hogan et al. in [9], we will propose algorithms to analyze the noises in the Linked Data.

Acknowledgements. We would like to thank Jiahong Shi, Yunxia Sun, and Yuan Si who contribute to this work. Additionally, we gratefully acknowledge funding from the National Science Foundation of China under grants 60873153, 60803061, and 61170165.

References

1. Auer, S., Lehmann, J.: Creating knowledge out of interlinked data. Semantic Web 1(1), 97–104 (2010)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)
3. Cimiano, P.: Ontology learning and population from text: algorithms, evaluation and applications. Springer, Heidelberg (2006)
4. d'Amato, C., Fanizzi, N., Esposito, F.: Inductive learning for the semantic web: What does it buy? Semantic Web 1(1-2), 53–59 (2010)
5. Domingos, P., Lowd, D.: Markov Logic: An interface layer for artificial intelligence. Synthesis Lectures on Artificial Intelligence and Machine Learning 3(1), 1–155 (2009)

6. Esposito, F., Fanizzi, N., Iannone, L., Palmisano, I., Semeraro, G.: Knowledge-intensive induction of terminologies from metadata. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 441–455. Springer, Heidelberg (2004)
7. Fanizzi, N., d'Amato, C., Esposito, F.: DL-FOIL concept learning in description logics. In: Železný, F., Lavrač, N. (eds.) ILP 2008. LNCS (LNAI), vol. 5194, pp. 107–121. Springer, Heidelberg (2008)
8. Getoor, L., Taskar, B.: Introduction to statistical relational learning. The MIT Press (2007)
9. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: 3rd International Workshop on Linked Data on the Web (LDOW 2010), in conjunction with 19th International World Wide Web Conference, CEUR (2010)
10. Ji, Q., Gao, Z., Huang, Z.: Reasoning with noisy semantic data. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol. 6644, pp. 497–502. Springer, Heidelberg (2011)
11. Lehmann, J.: Dl-learner: Learning concepts in description logics. The Journal of Machine Learning Research 10, 2639–2642 (2009)
12. Lehmann, J., Bühmann, L.: AutoSPARQL: Let users query your knowledge base. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 63–79. Springer, Heidelberg (2011)
13. Lehmann, J., Hitzler, P.: A refinement operator based learning algorithm for the ALC description logic. In: Blockeel, H., Ramon, J., Shavlik, J., Tadepalli, P. (eds.) ILP 2007. LNCS (LNAI), vol. 4894, pp. 147–160. Springer, Heidelberg (2008)
14. Lehmann, J., Hitzler, P.: Concept learning in description logics using refinement operators. Machine Learning 78, 203–250 (2010)
15. Lisi, F.A.: Building rules on top of ontologies for the semantic web with inductive logic programming. Theory and Practice of Logic Programming 8(03), 271–300 (2008)
16. Lisi, F.A.: Inductive logic programming in databases: From Datalog to DL+ log. Theory and Practice of Logic Programming 10(03), 331–359 (2010)
17. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems 16(2), 72–79 (2001)
18. Mitchell, T.: Machine learning. McGraw-Hill, New York (1997)
19. Niepert, M., Noessner, J., Stuckenschmidt, H.: Log-linear description logics. In: IJCAI, pp. 2153–2158 (2011)
20. Völker, J., Niepert, M.: Statistical schema induction. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 124–138. Springer, Heidelberg (2011)
21. Vrandecic, D., Krötzsch, M., Rudolph, S., Lösch, U.: Leveraging non-lexical knowledge for the linked open data web. The Fifth RAFT 2010 The yearly bilingual publication on nonchalant research 5(1), 18–27 (2010)
22. Zhu, M., Gao, Z., Qi, G., Ji, Q.: DLP learning from uncertain data. Tsinghua Science & Technology 15(6), 650–656 (2010)