Modeling the semantics of data sources with graph neural networks

Anonymous Authors[1]

Abstract

Semantic models are fundamental to publish data into Knowledge Graphs (KGs), since they encode the precise meaning of data sources, through concepts and properties defined within reference ontologies. However, building semantic models requires significant manual effort and expertise. In this paper, we present a novel approach based on Graph Neural Networks (GNNs) to build semantic models of data sources. GNNs are trained on Linked Data (LD) graphs, which serve as background knowledge to automatically infer the semantic relations connecting the attributes of a data source. At the best of our knowledge, this is the first approach that employs GNNs to identify the semantic relations. We tested our approach on 15 target sources from the advertising domain (used in other studies in the literature), and compared its performance against two baselines and a technique largely used in the state of the art. The evaluation showed that our approach outperforms the state of the art in cases of data source with the largest amount of semantic relations defined in the ground truth.

1. Introduction

Knowledge Graphs (KGs) are labeled multi-graphs that encode information as facts in the form of semantic entities and relations, which are relevant to a specific domain. Publishing data into KGs is a complex and time-consuming process, that typically requires extracting and integrating information from heterogeneous sources. The practice of integrating information from diverse types of data sources, such as CSVs, XMLs, and JSONs implies the construction of a map between the attributes of the data source and the concepts and properties defined by one or more ontologies (Gangemi, 2005). This map is formalized as a directed graph called *semantic model*, whose leaf nodes represent the attributes of the original data source, while the other parent nodes and edges derive from the properties and relations described in the reference ontologies. In order to transform the data source to KG facts, a semantic model can be used to automatically define rules in different mapping languages, such as RML (Dimou et al., 2014), R2RML (Das et al., 2016), TARQL (Cyganiak, 2015), or JARQL (Schiavone et al., 2018). Although semantic models can speed up the process of building a Knowledge Graph, its construction is a time-intensive task, since it requires significant effort and domain expertise, due to the potential variety and specificity of the data sources involved (e.g., it can be data from the Web or from private data lakes). In addition, the automatic extraction of the intended meaning of the data is a challenging process, which involves two main tasks. The first task is the *semantic labeling*, whose goal is to annotate the attributes of the data source with semantic labels (or semantic types). The second task is the *semantic relation inferencing*, whose goal is to capture the relations between the data source attributes. In this paper, we present a novel approach based on Graph Neural Networks (GNNs) to automatically identify the relations which connect already-annotated data attributes. GNNs have become the standard framework (Dwivedi et al., 2020) to learn from data on graphs for a variety of purposes, i.e. node and link prediction. In our method, GNNs are trained on Linked Data (LD) (Heath & Bizer, 2011) graphs that contain semantic information and act as background knowledge to reconstruct the semantics of data sources: the intuition is that relations used by other people to semantically describe data in a domain are more likely to express the semantics of the target source in the same domain. To measure the performance of our approach, we compared the results achieved by our system against ground-truth semantic models defined by domain experts. Furthermore, the evaluation procedure shows that our approach outperforms the state of the art (Taheriyan et al., 2016b) in case of data sources with the largest amount of semantic relations, according to the ground-truth semantic models.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Related Work

Influential works in the field (Taheriyan et al., 2013) (Taheriyan et al., 2016a) (Taheriyan et al., 2016b) indicate that research efforts in semantic modeling focused so far mainly on the semantic labeling, while less attention has been given to the automatic inference of semantic relations. The motivation for this observed trend has to be found in the complexity of the second step: in fact, even when semantic labels are properly defined with human intervention, inferring the relations through an automatic mechanism is not trivial and it is still an open issue in research. In addition, in more complex - but not unusual - situations, semantic labels can be connected through multiple paths that include different sequences of ontology classes and properties. As a consequence, without explicit and additional background context, it is difficult to identify which paths - or in other words which semantic relations - define the actual meaning of the data. Following this direction, the most promising approaches exploit background LD graphs, which include a vast amount of meaningful information, that can be used to learn how different entities are related to each other. As demonstrated by the work of Taheriyan et al. (Taheriyan et al., 2016b), a background knowledge is helpful to select a path representing the correct semantic interpretation of the target source. We took inspiration from this work to develop a novel mechanism based on GNNs for inferring semantic relations between data source attributes. The most important difference between our approach and the work of Taheriyan et al. (Taheriyan et al., 2016b) is that the latter manually extracts graph patterns to represent semantic relations of different lengths. In our approach, instead, the GNNs automatically learn entity and property representations, encoding the local multi-graph structures available in the LD. These representations are then exploited to identify the correct semantic relations within the target data source.

3. Problem definition

The problem of modeling the semantics of a data source is defined as follows. Suppose we have a target data source ds, which includes a set of attributes $ds\{a_1, a_2, a_3, ...\}$, and an ontology O. The semantic model of ds is defined as $sm(ds)$, whose generation is based on two different steps. The first step is the semantic labeling, where each attribute of ds is labeled with a pair of an ontology class and a data property: $sl_1(a_1) = \langle c_{a_1}, p_{a_1} \rangle$. The second step is the inference of the semantic relations between these semantic labels, expressing the intended meaning of the data. In the simplest case, the relation between two classes of the semantic labels includes only an object property: $sr_1(sl_1, sl_2) = c_{a_1} \xrightarrow{p_{o1}} c_{a_2}$. In this case the length of the path is equal to 1. In most complex situations, the relation covers different ontology classes and properties $sr_1(sl_1, sl_2) = c_{a_1} \xrightarrow{p_{o2}} c_1 \xrightarrow{p_{o3}} c_{a_2}$. In this case the length of the path is equal to 2.

4. Methodology

The starting point of our method is a multi, directed, and weighted graph, called integration graph: $G_{int} = V_{int}, E_{int}$. G_{int} describes the combinatorial space of all plausible semantic relations within the target source. The initial version of G_{int} is created from already annotated data source attributes and the ontology O, following the approach described by (Knoblock et al., 2012). Identifying the correct semantic relations in G_{int} corresponds to the detection of the minimum spanning tree, also called *Steiner Tree* (Hwang & Richards, 1992), in G_{int}. Considering that the detection of the Steiner Tree is driven by the costs associated to E_{int}, the goal of our methodology is to update these costs, whose role is to encode the correct interpretation of the data. To assign these costs, we employ a GNNs architecture, which learns entity and property features of LD graph, representing the background knowledge. The "recursive neighborhood diffusion" (Dwivedi et al., 2020) to assign entity features is based on an extension of the Vanilla Graph ConvNets (GCNs) (Kipf & Welling, 2016) formulation called Relational Graph ConvNets (R-GCNs) (Schlichtkrull et al., 2018):

$$h_i^{l+1} = ReLU\left(U^l \sum_{e \epsilon E_{ij}} \frac{1}{deg_i} \sum_{j \epsilon V_{ij}} h_j^l + h_i^l\right) \quad (1)$$

$h_i^l \in R^{d^{(l)}}$ denotes the hidden state of the LD entity i in the l-th layer of the GNNs. V_i^j is the set of indices of the neighbors j of entity i under the LD property $e \in E$. U^l is the matrix of the network parameters. By stacking up several layers, it is possible to capture and encode the relations between LD entities across multiple steps.

The function to score the predicted facts is the well-known matrix factorization algorithm called DistMult (Yang et al., 2014):

$$f\,(s, p, o) = (h_i^L)^T \, R_{e_{i,j}} \, h_j^L \quad (2)$$

h_i^L is the state of the entity i, as output of the recursive neighborhood diffusion. The features of the edge e are associated to a diagonal matrix $R_{e_{i,j}} \in R^{d \times d}$. The training of GNNs are performed with negative sampling. For each training sample, a set of negative samples w is generated by randomly corrupting either s or o. The network is optimized so that the positive facts are scored higher than the negative ones. The predicted fact score is equal to:

$$\hat{y} = \sigma(f(s, p, o)) \quad (3)$$

The cross entropy loss associated to each predicted fact is computed as follows:

$$L = -\frac{1}{(1+w)|E|}\sum_{\hat{y}\epsilon\tau} y\log\hat{y} + (1-y)\log(1-\hat{y}) \quad (4)$$

E is a subset of the LD edges included in the training set, w is the number of negative samples. The features of s, p, and o are computed during the network optimization. Then, the features and the scoring function are employed to compute the score of unseen facts, resulting from each plausible semantic relation in the integration graph. Each plausible relations allows to create a set of mapping language rules. These rules can used to generate a set of candidate facts $\{(s, p, o), ..\}$ from the data included in the source ds. s and o are instances of the ontology classes (nodes in the integration graph) included in sr, while p is an ontology property (edge in the integration graph) included in sr. The score of the facts associated to each plausible relation is computed with equation 3. Considering this score computation, the cost of each edge of the integration graph is the following:

$$cost(p_i) = \frac{1}{\frac{1}{|\tau|}\sum_{s,p_i,o\in s_r}\sigma(f(s,p_i,o))} \quad (5)$$

On the basis of the edges cost, the minimum spanning tree which connect all semantic labels (*Steiner Tree*) is detected in order to compute the most plausible semantic model, which includes the correct semantic relations to define the precise meaning of the data.

5. Evaluation

Dataset: the dataset includes 15 target sources available in JSON format on the advertising domain (Taheriyan et al., 2016b). The domain ontology is an extension of Schema.org (Guha et al., 2016), which contains 736 classes and 1081 properties. To prepare the background LD for each target source the leave-one-out setting has been employed. In practice, if k is the number of sources in our dataset, the background LD assigned to each target source is created from the facts obtained by the other $k-1$ sources. In other words, each background LD includes facts which come from all the sources, except those obtained from the target source. Details on the dataset are available in Table 1.

Metrics: the performance of the GNNs is evaluated with the Mean Reciprocal Rank (MRR). The accuracy of a computed semantic model sm is measured in terms of precision and recall, by comparing it against a ground-truth semantic model sm_{gt}:

$$precision = \frac{rel(sm_{gt}) \cap rel(sm)}{rel(sm)} \quad (6)$$

Table 1. Details on target sources, background linked data, and ground truth semantic models

Sources	#attrs	Background LD		Ground-Truth SMs	
		#entities	#facts	#labels	#relations
alaskaslist	8	3396	6954	12	3
armslist	20	3396	6793	15	4
dallasguns	15	3379	6940	23	7
elpasoguntrader	8	3396	7044	13	4
floridagunclassifieds	16	3396	6904	23	6
floridaguntrader	10	3396	6774	15	4
gunsinternational	10	3396	6945	19	4
hawaiiguntrader	7	3396	7122	11	3
kyclassifieds	10	3396	6945	14	3
montanagunclassifieds	9	3396	7104	14	4
msguntrader	11	3375	7086	16	4
nextechclassifieds	20	3396	6198	32	11
shooterswap	11	3396	7041	15	3
tennesseegunexchange	14	3396	7104	21	6
theoutdoorstrader	12	3396	6784	18	5

$$recall = \frac{rel(sm_{gt}) \cap rel(sm)}{rel(sm_{gt})} \quad (7)$$

where $rel(sm)$ is the set of triples (u, v, e): e is an object property from the ontology class u to the ontology class v.

Results: Table 2 reports: (i) details on the number of facts included in the training set, the validation set, and the testing set respectively; (ii) the resulting MRR on the testing set.

To measure the effectiveness of the GNNs on our background linked data, we compared our results with the MRR values obtained by the GNNs on FB15-k237(Toutanova & Chen, 2015). These MRR values reported in literature (Schlichtkrull et al., 2018) are: (i) MRR Raw: 0.158; (ii) Hits@1: 0.153; (iii) Hits@3: 0.258. MRR values obtained on background LD (Raw and Hits@1) are higher than the MRR values obtained on FB15-k237, therefore the GNNs performed well on the evaluation dataset.

Table 3 reports the results in terms of precision and recall achieved by: (i) our approach (Semi in the Table); (ii) the approach of Taheriyan et al. (Taheriyan et al., 2016b)) (Tahe in the Table); (iii) the baseline exploiting only the frequency of semantic relations of length 1 (Occs in the Table); (iv) the baseline using the steiner tree performed on a weighted graph based on the ontology structure (Knoblock et al., 2012) (Stei in the Table).

Our approach always obtained a better accuracy in terms of precision and recall, compared to: (i) the baseline that captures the frequency of semantic relations of length 1; (ii) the baseline of the steiner tree built on the graph weighted according to the ontology structure. In this experiment we employed the dataset in which the Taheriyan et al. (Taheriyan

Table 2. Number of facts in the training, the validation, and the testing set and the MRR values obtained by the GNNs on each background linked data

Sources	Background LD - #Facts			Mean Reciprocal Rank (MRR)		
	Training	Validation	Testing	Raw	Hits@1	Hits@3
alaskaslist	6264	345	345	0.202556	0.171014	0.221739
armslist	6123	335	335	0.189313	0.156716	0.214925
dallasguns	6250	345	345	0.222723	0.201449	0.233333
elpasoguntrader	6344	350	350	0.175496	0.135714	0.198571
floridagunclassifieds	6214	345	345	0.213165	0.191304	0.224638
floridaguntrader	6104	335	335	0.207233	0.174627	0.229851
gunsinternational	6264	345	345	0.205095	0.188406	0.211594
hawaiiguntrader	6412	355	355	0.208059	0.180282	0.223944
kyclassifieds	6255	345	345	0.191376	0.163768	0.207246
montanagunclassifieds	6394	355	355	0.233740	0.212676	0.245070
msguntrader	6386	350	350	0.209148	0.188571	0.222857
nextechclassifieds	5588	305	305	0.204046	0.177049	0.216393
shooterswap	6341	350	350	0.226965	0.205714	0.241429
tennesseegunexchange	3694	355	355	0.203350	0.180282	0.214085
theoutdoorstrader	6114	335	335	0.185680	0.159701	0.205970

Table 3. Results of the semantic relation inference in terms of precision and recall

Sources	Precision				Recall			
	Semi	Tahe	Occs	Stei	Semi	Tahe	Occs	Stei
alaskaslist	**1**	1	0.667	0	**1**	1	0.667	0
armslist	**0.750**	0.750	0.500	0	**0.750**	0.750	0.500	0
dallasguns	**0.667**	0.570	0.500	0	**0.570**	0.570	0.428	0
elpasoguntrader	0.500	1	0.500	0.250	0.500	0.750	0.500	0.250
floridagunclassifieds	**0.833**	0.800	0.167	0	**0.833**	0.670	0.167	0
floridaguntrader	**1**	1	0.750	0	**1**	1	0.750	0
gunsinternational	**0.750**	0.600	0.250	0	**0.750**	0.750	0.250	0
hawaiiguntrader	**1**	1	1	0	**1**	1	1	0
kyclassifieds	**1**	1	0.333	0.333	**1**	1	0.333	0.333
montanagunclassifieds	0.750	1	0.500	0	0.750	1	0.500	0
msguntrader	**0.670**	0.670	0.667	0	**0.500**	0.500	0.500	0
nextechclassifieds	0.454	1	0.182	0	**0.454**	0.360	0.182	0
shooterswap	**1**	0.750	1	0	**1**	1	1	0
tennesseegunexchange	0.667	1	0.500	0.167	0.667	1	0.500	0.167
theoutdoorstrader	0.800	0.830	0.200	0.200	0.800	1	0.200	0.200

et al., 2016b) approach obtained the best results. The results show that our approach outperforms the state of the art in case of the following data sources: “dallasguns”, “floridagunclassifieds”, “gunsinternational”, and “shooterswap”. These sources have the most complex structure in terms of number of semantic labels and semantic relations in the ground-truth semantic models (see Table 1 for more details). On the other side, the performance in terms of precision drops in presence of many data attributes within sources that are characterized by the same semantic type (see “elpasoguntrader” and “nextechclassifieds”). For instance, the “nextechclassifieds” source includes 5 different attributes that are labeled with the ontology class “schema:Offer”. According to the ground-truth semantic model of this source, the first attribute is linked to the other 4 attributes with the same object property. Nevertheless, this type of graph structure represents an anomaly because it never appears in the background knowledge of “nextechclassifieds”. We believe that including in the background LD analogous graph structures the performance should increase.

6. Conclusion

We proposed a novel GNNs-based model for automatically building semantic models of data sources. Our proposed approach achieves results comparable with the state-of-the-art method in the field. In the future, we would like to investigate more effective GNNs architectures to learn graph structures available in the background LD, to improve the accuracy of the computed semantic models.

References

Cyganiak, R. Tarql (sparql for tables): Turn csv into rdf using sparql syntax. Technical report, Technical report, Jan. 2015. http://tarql. github. io, 2015.

Das, S., Sundara, S., and Cyganiak, R. R2rml: Rdb to rdf mapping language. w3c recommendation (2012), 2016.

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van de Walle, R. Rml: a generic language for integrated rdf mappings of heterogeneous data. 2014.

Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., and Bresson, X. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982*, 2020.

Gangemi, A. Ontology design patterns for semantic web content. In *International semantic web conference*, pp. 262–276. Springer, 2005.

Guha, R. V., Brickley, D., and Macbeth, S. Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

Heath, T. and Bizer, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

Hwang, F. K. and Richards, D. S. Steiner tree problems. *Networks*, 22(1):55–89, 1992.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv e-prints*, art. arXiv:1609.02907, Sep 2016.

Knoblock, C. A., Szekely, P., Ambite, J. L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., and Mallick, P. Semi-automatically mapping structured sources into the semantic web. In *Extended Semantic Web Conference*, pp. 375–390. Springer, 2012.

Schiavone, L., Morando, F., Allavena, D., and Bevilacqua, G. Library data integration: the cobis linked open data project and portal. In *Italian Research Conference on Digital Libraries*, pp. 15–22. Springer, 2018.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pp. 593–607. Springer, 2018.

Taheriyan, M., Knoblock, C. A., Szekely, P., and Ambite, J. L. A graph-based approach to learn semantic descriptions of data sources. In *International Semantic Web Conference*, pp. 607–623. Springer, 2013.

Taheriyan, M., Knoblock, C. A., Szekely, P., and Ambite, J. L. Learning the semantics of structured data sources. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37:152–169, 2016a.

Taheriyan, M., Knoblock, C. A., Szekely, P., and Ambite, J. L. Leveraging linked data to discover semantic relations within data sources. In *International Semantic Web Conference*, pp. 549–565. Springer, 2016b.

Toutanova, K. and Chen, D. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.