RDF SUMMARIZATION

Luis Vives [1] César Beltran [2] Richard Chbeir [3]

[1−2]Pontificia Universidad Católica del Perú (PUCP)
[3]Université de Pau et des Pays de l'Adour (UPPA)

October 30, 2018

1. RDF previous
2. RDF definition
3. Advantages of RDF
4. Problems in RDF representation
5. RDF summarization
6. Advantages of RDF summarization
7. Our Proposal: RDF graph summarization with RNN

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

RDF previous

RDF was original writeen by Tim Bray in 1998 an update by Dan Brickley in 2001[15].

The Resource Description Framework (RDF) is a lenguage for representing information about resources in the World Wide Web[15].

RDF descriptions, often contain redundancies, and could be generated differently even when describing the same resources, which would have a negative impact on various RDF-based applications (e.g.,RDF storage, processing time, loading time, similarity measuring, mapping, alignment, and versioning)[13].

RDF definition

RDF graph: set of triples

Figure 2.1: Example of RDF triple

There can be three kinds of nodes: IRIs, literals, and blank nodes.

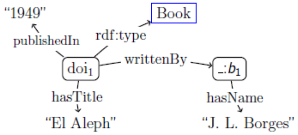
- **IRI** (Internationalized Resource Identifier) refers to a resource (the referent).

  Literal denotes resources which have an associated value for example, an integer or string value.

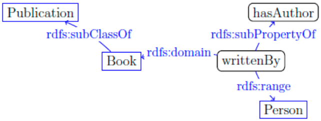
  Blank nodes are local identifiers which do not identify specific resources.

RDF definition

RDF data graph and RDF schema graph:



(a) RDF graph



(b) RDF Shema (RDFS) graph

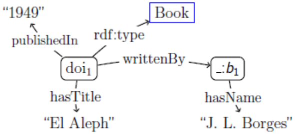
Figure 3.1: Example of RDF graph and RDFS graph

RDF definition

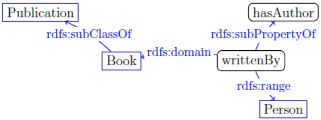
RDF Statement	Triple	Shorthand
Class Assertion	(s,rdf:type,o)	(s,t,o)
Property assertion	(s,p,o) with p ≠ rdf:type	(s,p,o)

RDFS Statement	Triple	Shorthand
SubClass	(s,rdf:subClassOf,o)	(s,$\prec_{sc}$,o)
Property assertion	(s,rdf:subPropertyOf,o)	(s,$\prec_{sp}$,o)
Domain typing	(s,rdf:domain,o)	(s,$\hookleftarrow_{d}$,o)
Range typing	(s,rdf:range,o)	(s,$\hookleftarrow_{d}$,o)

Table 1: RDF & RDFS statements



(a) RDF graph



(b) RDF Shema (RDFS) graph

Figure 4.1: Example of RDF graph and RDFS graph

Advantages of RDF

- Efforts to generate knowledge base of RDF scalable, fast, secure[10].
- Efforts to incorporate data quality metrics in RDF queries [16][3][2].
- Efforts to process multiple RDF networks in parallel [4][8].
- Efforts to consult information in RDF based on data patterns, keywords[8].

Problems in RDF representation

1. RDF datasets are growing constantly.
2. Minimum Constraints for RDF data make it irregular, difficult to comprehend and visualize, this can cause problems for information extraction, processing, and analysis.
3. RDFs have been designed as a query standard based on explicit and implicit data.

Many authors propose use **RDF summarization** approach to solve problem item 2 y 3

RDF summarization

RDF graph are often large and varied, produced in a variety of contexts such as social networks, medical data, scientific data, etc. [9]. The large amount of data contained in the RDF is often too expensive to perform queries to acquire information. The RDF summary refers to the process of extracting concise but

significant summaries of RDF Knowledge Bases (KBs) that represent as close as possible the actual contents of the KB, both in terms of structure and data [17].

RDF summary requerimients

We are interested in extracting a summary graph, having the following characteristics:

- The summary is a RDF graph: The summary graph should be a RDF graph itself.

  The size of the Summary: The volume of a graph is the numbers of its edges and nodes.

  Thus the summary graph should:

  - Be smaller than the original RDF graph.

    Contain all the important information.

    Report the most representative nodes (classes) and edges (properties).

Interest in research towards the quality of the summary[7][17]

Main Approaches for RDF Summarization[17]

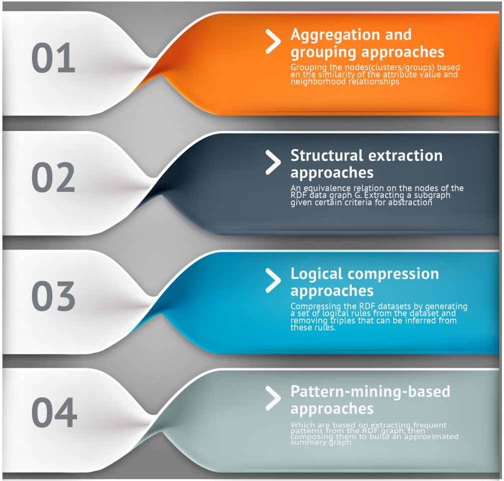


Figure 7.1: Main Approaches for RDF Summarization[17]

Advantages of RDF summarization

- Efforts to manage heterogeneous and homogeneous RDF networks[1].
- Efforts to obtain quality summaries[7].
- Efforts to reduce the loss of information in RDF summaries.
- Efforts to interpret explicit and implicit information in RDF summaries[11].

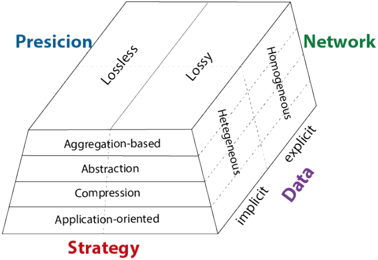


Figure 8.1: Advantages of RDF summarization

Our Proposal: RDF graph summarization with Recurrent Neural Network (RNN)

In the paper: **"Modeling Relational Data with Graph Convolutional Networks"**[12] exist some impressions:

- General Fourier transform scales poorly with size of data so we need relaxations, normally use for image processing, where actual spatial convolutions are easy to compute.
- At all levels in this network, the filters are limited to 3x3 in size and are also essentially fixed to be the same kernel across all layers and all units in entire network

In the paper **"Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering"?**], the authors try to solve the above problems including higher order Chebyshev polynomials in the approximation.

For satisfy the stationarity, locality, compositionality assumptions, we propose use **"Recurrent Neural Networks"**

RDF summarization with Recurrent Neural Network (RNN)

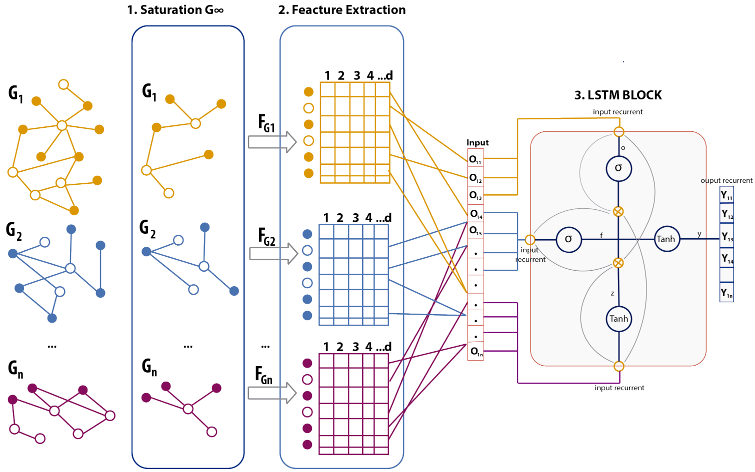


Figure 9.1: RDF graph summarization with RNN

Saturation

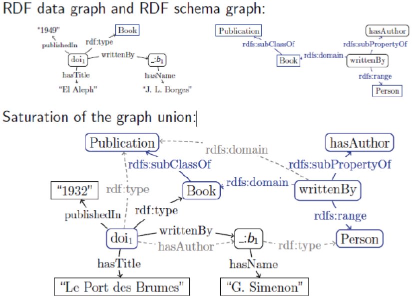


Figure 9.2: Saturation of the graph union

Rule	Entailment rule
rdfs2	$(p, \hookleftarrow_d, o), (s_1, p, o_1) \rightarrow (s_1, \tau, o)$
rdfs3	$(p, \hookrightarrow_r, o), (s_1, p, o_1) \rightarrow (o_1, \tau, o)$
rdfs5	$(p_1, \preceq_{sp}, p_2), (p_2, \preceq_{sp}, p_3) \rightarrow (p_1, \preceq_{sp}, p_3)$
rdfs7	$(p_1, \preceq_{sp}, p_2), (s, p_1, o) \rightarrow (s, p_2, o)$
rdfs9	$(s, \preceq_{sc}, o), (s_1, \tau, s) \rightarrow (s_1, \tau, o)$
rdfs11	$(s, \preceq_{sc}, o), (o, \preceq_{sc}, o_1) \rightarrow (s, \preceq_{sc}, o_1)$
ext1	$(p, \hookleftarrow_d, o), (o, \preceq_{sc}, o_1) \rightarrow (p, \hookleftarrow_d, o_1)$
ext2	$(p, \hookrightarrow_r, o), (o, \preceq_{sc}, o_1) \rightarrow (p, \hookrightarrow_r, o_1)$
ext3	$(p, \preceq_{sp}, p_1), (p_1, \hookleftarrow_d, o) \rightarrow (p, \hookleftarrow_d, o)$
ext4	$(p, \preceq_{sp}, p_1), (p_1, \hookrightarrow_r, o) \rightarrow (p, \hookrightarrow_r, o)$

Figure 9.3: Sample RDF entailment rules[5]

Other authors[14][12] assign labels to entity types, for example: “John Doe”, “London”, “England”, and “1967-01-10” can be mapped to “PERSON”, “CITY”, “COUNTRY”, and “DATE”

Feature extraction

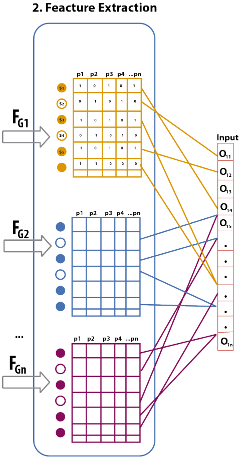


Figure 9.4: Feacture Extraction

Other technique for feacture extraction[1]

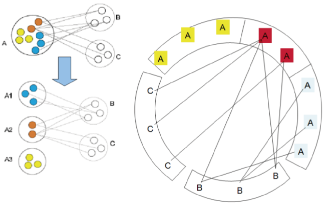


Figure 9.5: Example of aggregation using K-Snap

Recurrent Neural Networks

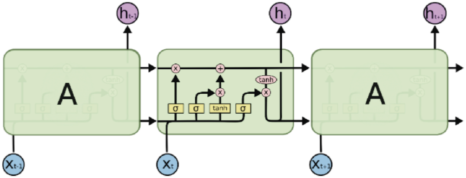


Figure 9.6: Recurrent Neural Networks Architecture

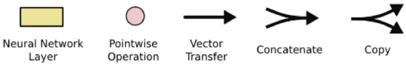


Figure 9.7: Recurrent Neural Networks Architecture

Recurrent neural networks address the problem reasoning about previous events in the nodes to inform later ones. They are networks with loops in them, allowing information to persist[6].

References I

[1] Aufaure, A. and Louati, M. A. (2012). Graph Aggregation : Application to Social Networks Graph Aggregation Algorithms.

[2] Bursztyn, D., Goasdoué, F., and Manolescu, I. (2014). Optimizing Reformulation-based Query Answering in RDF. *18th International Conference on Extending Database Technology (EDBT)*, pages 265–276.

[3] Cheng, G., Ji, F., Luo, S., Ge, W., and Qu, Y. (2012). BipRank: Ranking and summarizing RDF vocabulary descriptions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7185 LNCS:226–241.

[4] Doel, T., Shakir, D. I., Pratt, R., Aertsen, M., Moggridge, J., Bellon, E., David, A. L., Deprest, J., Vercauteren, T., and Ourselin, S. (2017). GIFT-Cloud: A data sharing and collaboration platform for medical imaging research. *Computer Methods and Programs in Biomedicine*, 139:181–190.

[5] Hassad, S. E. (2017). Learning Commonalities in RDF. pages 502–517.

[6] Jozefowicz, R. and Com, I. G. (2015). An Empirical Exploration of Recurrent Network Architectures. 37.

[7] Khatchadourian, S. and Consens, M. P. (2010). ExpLOD : Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. pages 272–287.

[8] Kushwaha, N., Singh, B., Mahule, R., and Vyas, O. P. (2015). Keyword prediction with ARM on bibliographic RDF data. *Procedia Computer Science*, 50:490–495.

[9] Manolescu, I. (2018). Structural Summarization of Semantic Graphs Ioana Manolescu To cite this version : HAL Id : hal-01808737 Structural Summarization of Semantic Graphs.

[10] Ren, X., Curé, O., Naacke, H., and Xiao, G. (2018). BigSR: an empirical study of real-time expressive RDF stream reasoning on modern Big Data platforms. pages 1–16.

[11] Ristoski, P. and Paulheim, H. (2016). RDF2Vec : RDF Graph Embeddings for Data Mining. page 30.

[12] Schlichtkrull, M., Kipf, T. N., Bloem, P., Titov, I., and Welling, M. (2017). Modeling Relational Data with Graph Convolutional Networks. (1).

References II

[13] Ticona, R., Tekli, J., Dongo, I., Guzman, R., and Chbeir, R. (2017). Toward RDF Normalization.

[14] Trisedya, B. D., Qi, J., Zhang, R., and Wang, W. (2018). GTR-LSTM: A Triple Encoder for Sentence Generation from RDF Data. *Proceedings of ACL*, pages 1627–1637.

[15] w3c (2004). Resource description framework (rdf):concepts and abstract syntax. [Web; accedido el 06-10-2018].

[16] Zneika, M., Lucchese, C., Vodislav, D., and Kotzinos, D. (2016). Summarizing Linked Data RDF Graphs Using Approximate Graph Pattern Mining. *Proc. 19th International Conference on Extending Database Technology*, pages 684–685.

[17] Zneika, M., Vodislav, D., Kotzinos, D., Metrics, Q., Rdf, F., and Summarization, G. (2018). Quality Metrics For RDF Graph Summarization.

Thank You.

Luis Vives [1], César Beltran [2], Richard Chbeir [3]