Alignment aware Linked Data Compression

Amit Krishna Joshi, Pascal Hitzler, and Guozhu Dong

Wright State University, Dayton, OH, U.S.A.
{joshi.35, pascal.hitzler, guozhu.dong}@wright.edu

Abstract. The success of linked data has resulted in a large amount of data being generated in a standard RDF format. Various techniques have been explored to generate a compressed version of RDF datasets for archival and transmission purpose. However, these compression techniques are designed to compress a given dataset without using any external knowledge, either through a compact representation or removal of semantic redundancies present in the dataset. In this paper, we introduce a novel approach to compress RDF datasets by exploiting alignments present across various datasets at both instance and schema level. Our system generates lossy compression based on the confidence value of relation between the terms. We also present a comprehensive evaluation of the approach by using reference alignment from OAEI.

1 Introduction

Linked data has experienced accelerated growth in recent years due to its interlinking ability across disparate sources, made possible via machine processable non-proprietary RDF data. Today, large number of organizations, including governments and news providers, publish data in RDF format, inviting developers to build useful applications through re-use and integration of data. This has led to tremendous growth in the amount of RDF data being published on the web. Although the growth of RDF data can be viewed as a positive sign for semantic web initiatives, it also causes performance bottlenecks for RDF data management systems that store and provide access to it [8]. As such, the need for compressing structured data is becoming increasingly important.

Various RDF representations and compression techniques have been developed to reduce the size of RDF data for storage and transmission. Representation like N3, Turtle and RDF/JSON offer compactness while maintaining readability by reducing verbosity of original RDF/XML format. Earlier RDF compression studies [3, 4, 18] focused on dictionary encoding and RDF serialization techniques. [4] proposed a new compact representation format called Header-Dictionary-Triples (HDT) that takes advantage of skewed data in large RDF graphs. [9] introduced the notion of a lean graph which is obtained by eliminating triples with blank nodes that specify redundant information. [17] and [14] studied the problem of redundancy elimination on RDF graphs in the presence of rules and constraints. [12] introduced rule based compression technique that exploits the semantic redundancies present in large RDF graph by mining frequent patterns and removing triples that can be identified by association rules.

These techniques perform compression by identifying syntactic and semantic redundancies in a dataset but do not exploit alignments present across various datasets. In this paper, we propose a novel technique to incorporate alignments for compressing datasets. This is the first study investigating lossy RDF compression based on alignments and application context.
This work was supported by the National Science Foundation under award 1017225 III: TROn (Tractable Reasoning with Ontologies).

2 Motivation

Ontology Alignment [16] refers to the task of finding correspondences between ontologies. It's a widely explored topic and numerous applications have been developed that perform the task of ontology alignment and mapping for schemas and instances. It finds a number of use cases including schema integration, data integration, ontology evolution, agent communication and query answering on the web [15, 13].

In this study, we utilize ontology alignments for compressing RDF datasets. Below, we list a few properties that could be exploited while compressing multiple datasets.

Schema heterogeneity and Alignment:
Linked datasets cater to different domains, and thus require different modeling schemas. Even when datasets belong to the same domain, they could be modeled differently depending on the creator. For instance, Jamendo[1], and BBC Music[2] both belong to music domain but they use different ontologies[3,4]. Different ontologies, whether belonging to the same domain or not, often share some similarities and the terms can be aligned. Based on the resulting alignment, individual datasets can be rewritten using fewer schema terms before further processing. Many studies have been focused on schema alignment using various approaches such as sense clustering [6], instance similarity [10, 19] and structural/lexical similarities[11]. Factforge [1] uses an upper level PROTON[5] as a reference layer and has more than 500+ mapping across various datasets.
Datasets rewritten using a set of mapping terms lead to increased occurrences of same terms, resulting in a better compression.

Entity Co-reference and Linking:

The purpose of entity co-reference is to determine if different resources refer to the same real world entity. Often datasets have overlapping domains and tend to provide information about the same entity [5]. One of the approach include using similarity properties such as owl:sameAs or skos:exactMatch. For instance, LinkedMdb provides information about the Romeo & Juliet movie and provides direct reference to DBPedia using the owl:sameAs property.

[1] http://dbtune.org/jamendo/
[2] http://www.bbc.co.uk/music
[3] http://musicontology.com/
[4] http://www.bbc.co.uk/ontologies/bbc
[5] http://www.ontotext.com/proton-ontology/

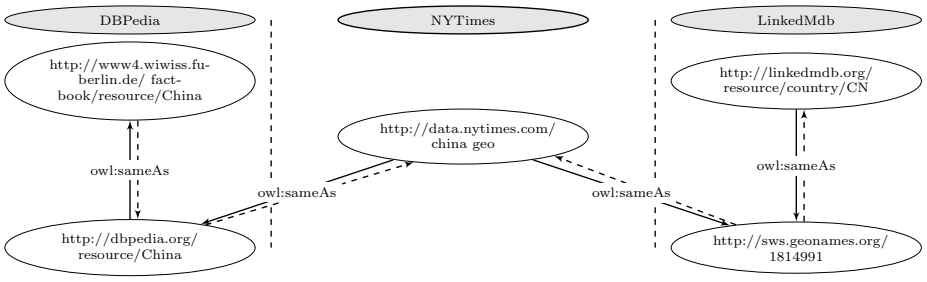

Fig. 1. LinkedMdb connects to DBPedia via NYTimes

However, there are cases where the two instances might not be directly connected but a path exists for such a co-reference as shown in Figure 1. Here, the Geonames resource for China is linked to the CIA Factbook concept and the DBPedia concept for China, using an owl:sameAs" link from the NYTimes dataset.

Varied Alignments: Alignment results vary greatly among different ontology matching systems (see [7]). Some of these work best for one set of ontologies while perform low in a different set of ontologies. The alignments can differ even when manually performed among group of experts for the same set of ontologies. For instance, conference track of OAEI provides the reference alignments[6] with a confidence score of 1 (signifying exact match) for all mappings within a collection of ontologies describing the domain of organizing conferences. On the contrary, [2] introduced a new version of the Conference reference alignment for OAEI that includes the varying confidence values reflecting expert disagreement on the matches.

3 Alignment aware Linked Data Compression

In this section, we elaborate on the internals of our compression system. The main task involves identification of alignments across various datasets. The alignments can be manual or generated using existing Ontology matching systems[7].

Given two ontologies O_i and O_j, we can compute multiple mappings between the ontology terms, t_i and t_j.
Alignment, μ is defined as $\mu =< t_i, t_j, r, s >$ where r denotes the relationship and $s \in [0, 1]$ is the confidence score that the relationship holds in the mapping.

Fig 5 represents the high level overview of our system. Given a set of input datasets, we first identify alignments present across these datasets. For this, we extract terms from each dataset and check for alignments with other participating datasets either manually or using automated ontology matching systems. It should be noted that the alignments can be in both schema and instance level. The set of alignments are then consolidated by performing mapping to a set of

[6] http://oaei.ontologymatching.org/2014/conference/data/reference-alignment.zip
[7] http://www.mkbergman.com/1769/50-ontology-mapping-and-alignment-tools/

master terms and pruning all mappings that have a confidence score below the threshold.

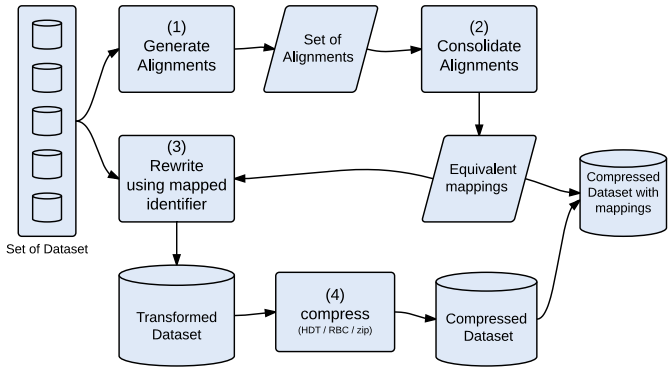

Fig. 2. Conceptual System Overview

The resulting unique set of mappings, together with the original datasets go through a transformation phase where all datasets are merged and the equivalent terms are replaced with *master* terms. Once the transformation is complete, the combined dataset is then passed to existing compression systems such as HDT, RBC and Zip to generate compressed dataset.

```
<http://ekaw#Regular_Paper>
-<http://cmt#PaperFullVersion>
-<http://confOf#Contribution>
-<http://iasted#Submission>
-<http://cmt#Paper>
-<http://ekaw#Regular_Paper>
-<http://edas#Paper>
-<http://confOf#Paper>
-<http://sigkdd#Paper>
-<http://conference#Paper>
-<http://ekaw#Paper>
```

```
<http://ekaw#Research_Topic>
-<http://cmt#SubjectArea>
-<http://edas#Topic>
-<http://confOf#Topic>
-<http://ekaw#Research_Topic>
-<http://conference#Topic>
```

Fig. 3. Sample grouping of equivalent terms for ekaw#Regular_Paper and ekaw#Research_Topic using OAEI reference alignment.

Algorithm for the consolidation of alignments is listed in Algorithm 1. Given a threshold and a set of alignments, mappings with confidence score less than a threshold are pruned and a set of master items is generated. Each master item maps to a group of equivalent ontology terms. These master items are later used to rewrite the dataset to replace ontology terms with corresponding master item. The alignments can contain both instance and schema terms. Fig.3 shows two master items and corresponding consolidated alignments.

4 Evaluation

For this paper, we built a prototype, LinkIt, in JAVA to test the validity of our approach. We experimented using reference alignments from OAEI.

Algorithm 1 Consolidation of Alignments

Require: A Alignment set and θ threshold for alignments

Pruning mappings with confidence lower than threshold value

Valid Mapping M as $\langle k, V\rangle \leftarrow \phi$
Term Mapping $G \leftarrow \phi$
Set $S \leftarrow \phi$
MasterItem Mapping $I \leftarrow \phi$
for each mapping, $< e1, e2, r, s >$ that occurs in A **do**
  if $r =' equivalence'$ and $s >= \theta$ **then**
    $M \leftarrow M \cup \langle e1, V \cup e1\rangle$ ▷ add a new valid mapping
    $M \leftarrow M \cup \langle e2, V \cup e2\rangle$
  end if
end for

Grouping equivalent terms

for all $\langle k, V\rangle$ in M **do**
  if $k \notin keys(G)$ and $k \notin S$ **then**
    $G \leftarrow G \cup \langle k, V_k\rangle$ ▷ mark this k as master item
    $S \leftarrow S \cup k$ ▷ mark this k as processed item
    for each $t \in V_k$ **do** ▷ group all items in V_k under k
    $G \leftarrow G \cup \langle k, V_t\rangle$ ▷ k maps to $V_k \cup V_t$
    $S \leftarrow S \cup t$ ▷ mark this t as processed item
    end for
  end if
end for

One to One mapping with master item

for each $(k, V) \in G$ **do**
  for each $v \in V$ **do**
    $I \leftarrow I \cup \langle v, k\rangle$ ▷ map to master item
  end for
end for

4.1 Dataset Generation

Since our primary purpose is to validate that RDF data can be compressed in presence of alignments, we need a set of ontologies, reference alignment for those ontologies and RDF data large enough to be tested. For the evaluation, we generated large size of synthetic RDF data using SyGENiA[8] tool and a set of Conference ontologies and the reference ontologies available from OAEI[9]. Given a set of queries and an ontology, SyGENiA tool can automatically generate a large number of individuals. The set of queries that we use for generating RDF data is available from[10]. In order to test the compression against dataset of varying size, we created multiple queries and generated eight different dataset. The size of evaluation dataset size is shown in Table 4.

[8] https://code.google.com/p/sygenia/
[9] http://oaei.ontologymatching.org/2014/
[10] http://bit.ly/1hgNsRv

	Dataset size(MB) created using query							
ontology	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Conference	113	261	257	123	195	213	113	727
confOf	107	152	149	77	137	129	98	546
iasted	84	161	157	74	129	108	84	670
sigkdd	98	158	146	92	137	126	88	390
cmt	67	149	140	79	97	99	56	658
edas	107	192	181	90	137	139	108	769
ekaw	94	181	177	63	146	147	92	704
Total	670	1254	1207	598	978	961	639	4464

Fig. 4. Dataset size for various set of queries.

4.2 Varied Alignments and Compression

We evaluated two versions of Conference reference alignment available from OAEI and [2]. These reference alignments include 16 ontologies related to the conference organization and they are based upon the actual conference series and corresponding web pages[11]. The mappings in the Conference:V1 are all set to be exact match. Figure 5 compares the distribution of valid mappings for various thresholds for both reference alignments. The number of mappings are generated after the consolidation of alignments. As expected, the number of mapping decreases with the increase of threshold in Conference:V2 reference alignment.

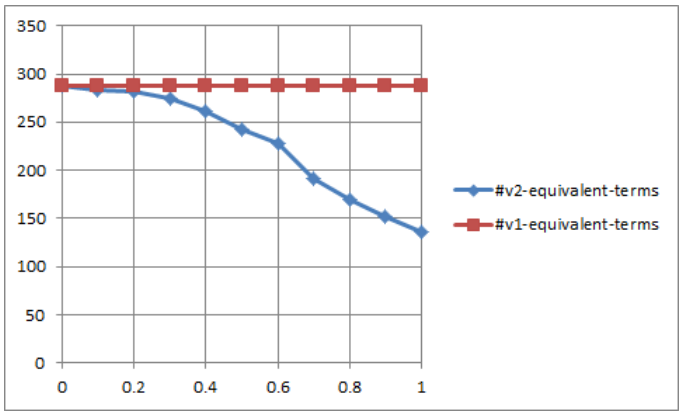

Fig. 5. Number of mappings at different thresholds for two versions of Conference reference alignments

Furthermore, for the same set of datasets, various ontology matching systems can produce different set of alignments. Fig.6 shows a comparison of various alignment systems with varying number of equivalent terms for same threshold, as seen in the results of OAEI[12]. The alignments are generated for the same set of ontologies used in Conference:V1 and Conference:V2 reference alignments.

[11] http://oaei.ontologymatching.org/2014/conference/index.html
[12] http://oaei.ontologymatching.org/2014/conference/eval.html

	Alignment System					
threshold	AML	Xmap	RSDLWB	OMReasoner	MaasMtch	LogMap
1	9	134	162	154	219	194
0.9	146	134	162	154	482	204
0.8	170	143	162	154	532	213
0.7	194	145	162	154	532	218
0.6	213	145	162	154	532	225
0.5	220	146	162	154	532	230
0.4	220	148	162	154	532	238
0.3	220	148	162	154	532	239
0.2	220	148	162	154	532	240
0.1	220	148	162	154	532	240

Fig. 6. Comparison of various automated alignment systems demonstrating varying number of equivalent terms for same threshold

As seen in Fig. 6, some alignment systems such as RSDLWB and OMReasoner generate all alignments with a confidence score of 1, while others like LogMap and XMap generate alignments with varying confidence score.

Since the alignment is not one to one, we cannot recover the original data once compressed and hence the compression is lossy.

The evaluation result for varying alignments is shown in Fig.7 for one of the datasets which has original size of 670MB. The compressed size can be compared against the output resulting from HDT alone which is 56MB.

AlignmentSystem	Compressed size (MB)
V1	51
V2	53
AML	53
Logmap	51
OmReasoner	52
Maasmtch	51
rsdlwb	53
xmap	53

Fig. 7. Compressed size (in MB) against original size of 670MB

5 Conclusion

In this paper, we have introduced a novel compression technique that exploits alignments present across various datasets at schema and instance level. We have explored lossy RDF compression, the area which has barely been researched in the semantic web field. The system extracts all mappings with confidence score greater or equal to the threshold and group them using single identifiers. Hence, our approach is flexible enough to cut-off alignments based on threshold that are context dependent.

In the future, our research can be directed towards finding applications of lossy RDF compression. We will also explore the effect of alignments in compressing RDF streams.

References

1. Bishop, B., Kiryakov, A., Ognyanov, D., Peikov, I., Tashev, Z., Velkov, R.: Factforge: A fast track to the web of data. Semantic Web 2(2), 157–166 (2011)
2. Cheatham, M., Hitzler, P.: Conference v2. 0: An uncertain version of the OAEI Conference benchmark. In: The Semantic Web–ISWC 2014, pp. 33–48. Springer (2014)
3. Fernández, J.D., Gutierrez, C., Martínez-Prieto, M.A.: RDF compression: basic approaches. In: Proceedings of the 19th international conference on World wide web. pp. 1091–1092. ACM (2010)
4. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange. Web Semantics: Science, Services and Agents on the World Wide Web 19, 22–41 (2013)
5. Glaser, H., Jaffri, A., Millard, I.: Managing co-reference on the semantic web (2009)
6. Gracia, J., d'Aquin, M., Mena, E.: Large scale integration of senses for the semantic web. In: Proceedings of the 18th international conference on World wide web. pp. 611–620. ACM (2009)
7. Grau, B.C., Dragisic, Z., Eckert, K., Euzenat, J., Ferrara, A., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Kempf, A.O., Lambrix, P., et al.: Results of the Ontology Alignment Evaluation Initiative 2013. In: Proc. 8th ISWC workshop on ontology matching (OM). pp. 61–100. No commercial editor. (2013)
8. Huang, J., Abadi, D.J., Ren, K.: Scalable SPARQL querying of large rdf graphs. Proceedings of the VLDB Endowment 4(11), 1123–1134 (2011)
9. Iannone, L., Palmisano, I., Redavid, D.: Optimizing RDF storage removing redundancies: An Algorithm. In: Innovations in Applied Artificial Intelligence, pp. 732–742. Springer (2005)
10. Isaac, A., Van Der Meij, L., Schlobach, S., Wang, S.: An empirical study of instance-based ontology matching. Springer (2007)
11. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 235–251 (2009)
12. Joshi, A.K., Hitzler, P., Dong, G.: Logical linked data compression. In: The Semantic Web: Semantics and Big Data. pp. 170–184. Springer (2013)
13. Joshi, A.K., Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P., Damova, M.: Alignment-based querying of linked open data. In: On the Move to Meaningful Internet Systems: OTM 2012, pp. 807–824. Springer (2012)
14. Meier, M.: Towards rule-based minimization of RDF graphs under constraints. In: Web Reasoning and Rule Systems, pp. 89–103. Springer (2008)
15. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record 33(4), 65–70 (2004)
16. Noy, N., Stuckenschmidt, H.: Ontology alignment: An annotated bibliography. Semantic Interoperability and Integration 4391 (2005)
17. Pichler, R., Polleres, A., Skritek, S., Woltran, S.: Redundancy elimination on RDF graphs in the presence of rules, constraints, and queries. In: Web Reasoning and Rule Systems, pp. 133–148. Springer (2010)
18. Urbani, J., Maassen, J., Drost, N., Seinstra, F., Bal, H.: Scalable RDF data compression with MapReduce. Concurrency and Computation: Practice and Experience 25(1), 24–39 (2013)
19. Wang, S., Englebienne, G., Schlobach, S.: Learning concept mappings from instance similarity. Springer (2008)