TRTML - A Tripleset Recommendation Tool based on Supervised Learning Algorithms

Alexander Arturo Mera Caraballo[1], Narciso Moura Arruda Júnior[2], Bernardo Pereira Nunes[1], Giseli Rabello Lopes[1], Marco Antonio Casanova[1]

[1] Department of Informatics, PUC-Rio, Rio de Janeiro/RJ – Brazil
{acaraballo,bnunes,grlopes,casanova}@inf.puc-rio.br
[2] Computer Science Department, UFC, Fortaleza/CE – Brazil
narciso@lia.ufc.br

Abstract. *The Linked Data initiative promotes the publication of interlinked RDF triplesets, thereby creating a global scale data space. However, to enable the creation of such data space, the publisher of a tripleset t must be aware of other triplesets that he can interlink with t. Towards this end, this paper describes a Web-based application, called TRTML, that explores metadata available in Linked Data catalogs to provide data publishers with recommendations of related triplesets. TRTML combines supervised learning algorithms and link prediction measures to provide recommendations. The evaluation of the tool adopted as ground truth a set of links obtained from metadata stored in the DataHub catalog. The high precision and recall results demonstrate the usefulness of TRTML.*

Keywords: Linked Data, Recommender Systems, Link Prediction, Machine Learning

1 Introduction

Over the past years, data publishers have been encouraged to publish their data following the Linked Data principles to facilitate data sharing, data reuse and enhance (semantic) interoperability on the Web [1, 2]. The main idea behind Linked Data is to connect resources across triplesets and, thereby, facilitate the discovery of related resources [3], the integration of data sources [4] and the enrichment of datasets [5].

However, with the steady growth of the number of triplesets published on the Web and the lack of tools to recommend and interlink related triplesets, most data publishers rely on a few reference data sources, such as DBpedia, Freebase and Geonames, to interlink their triplesets, leaving out other potentially related triplesets. As an attempt to assist data publishers in the process of tripleset interlinking, the Linked Data community created metadata catalogs describing triplesets (e.g. DataHub). Despite the existence of such catalogs, the arduous and laborious task of searching for related triplesets remains. Furthermore, a recent research [6] shows that metadata catalogs are often outdated and miss relevant information, further hindering the process of tripleset interlinking.

Thus, in this paper, we describe a Web-based application, called *TRTML*, that provides recommendations of triplesets related to a given tripleset. *TRTML* relies on supervised algorithms (such as Multilayer Perceptron, Decision Trees - J48 and Support Vector Machines) and link prediction measures (such as Common Neighbors, Jaccard coefficient, Preferential Attachment and Resource Allocation) that explore a set of features (e.g. vocabularies, classes and properties) available for the triplesets in data catalogs. In particular, the supervised learning algorithms are responsible for determining the best set of features for the recommendation task.

To evaluate the tool, we adopted as ground truth a set L of links obtained from metadata stored in the DataHub catalog. Briefly, we removed some of the links in L and evaluated, in terms of precision, recall and F-measure, how many of the removed links the *TRTML* tool was able to find. The experiments show that *TRTML* achieves an F-measure of 78%.

The rest of this paper is organized as follows. Section 2 presents an overview of the TRTML tool along with the supervised learning algorithms, link prediction measures and features used. Section 3 describes the evaluation setup and the results achieved. Finally, Section 4 summarizes the contributions and results.

2 Tripleset Recommendation Approach

Let $D = \{d_1, ..., d_n\}$ be a set of triplesets considered in the recommendation process and t be the tripleset one wants to receive recommendations for interlinking. Instead of providing a restricted list of recommendations, we define the task of recommending triplesets to be interlinked with t as a task of ranking triplesets d_i in D according to the estimated probability that one can define links between resources of t and d_i. To generate the rankings, we explore an approach that combines link prediction measures and machine learning techniques.

Link prediction measures. The approach uses link prediction measures to estimate the likelihood of the existence of a link between triplesets. To estimate the measures, we construct a bipartite graph $G = (D, F, E)$ consisting of two disjoints sets of nodes representing triplesets D and features F. The set of edges E represents the association between the triplesets and their features. The set of features of a tripleset t, F_t, correspond to the vocabularies, classes or properties extracted from the VoID descriptions defined in t. The tool implements four of the traditional link prediction measures, summarized in Table 1, which demonstrated good performance in previous works [7, 8].

Supervised learning algorithms. The approach uses supervised learning algorithms to learn if a pair of triplesets can be interlinked, using as training set the existing links between triplesets. Specifically, we build a J48 decision tree (Quinlan's C4.5 implementation), where the nodes represent the measures reported in Table 1, estimated using different feature sets (vocabularies, classes or properties). The leaf nodes represent the values of a binary class such that, given two triplesets (t, d_i), 1 represents that d_i can be recommended to t and 0

Table 1: Link prediction measures

Measure	Equation
Common Neighbors	$CN_{t,d_i} = \lvert F_t \cap F_{d_i} \rvert$
Jaccard coefficient	$Jaccard_{t,d_i} = \frac{\lvert F_t \cap F_{d_i} \rvert}{\lvert F_t \cup F_{d_i} \rvert}$
Preferential Attachment	$PA_{t,d_i} = \lvert F_t \rvert . \lvert F_{d_i} \rvert$
Resource Allocation	$RA_{t,d_i} = \sum_{f_j \in F_t \cap F_{d_i}} \frac{1}{\lvert D_{f_j} \rvert}$

Where:

- F_{d_i} is the feature set of tripleset d_i (direct neighbors of d_i in G);
- D_{f_j} is the set of triplesets having feature f_j (direct neighbors of f_j in G).

denotes that d_i is not a good candidate to be recommended to t. The advantage of decision tree classifiers over other supervised learning algorithms is that they produce an interpretable model that allows users to understand how to classify new instances.

TRTML Overview. Suppose that a user is working on a tripleset t and that he wants to discover one or more triplesets d_i such that t can be interlinked with d_i. He then uses the tool to obtain tripleset recommendations. First, the tool builds a classifier over the set of VoID descriptions, obtained from the DataHub catalog. Then, the user defines the rest of the input data the tool requires: (i) he selects the serialization format of the VoID descriptor (TURTLE, RDF/XML or N-TRIPLE N3); and (ii) uploads a VoID descriptor V_t for t from which the tool extracts the feature set F_t by analyzing the `void:vocabulary`, `void:class` and `void:property` occurring in V_t. Finally, the tool applies the classifier, using F_t, and outputs a ranked list of triplesets, sorted by the estimated probability of creating links with t.

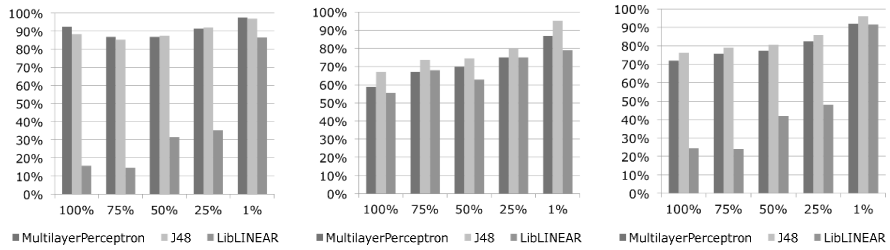
The tool is available at `http://web.ccead.puc-rio.br:8080/Uncover/ml/`.

3 Experimental evaluation

Triplesets. We based the experiments on the VoID descriptions stored in the DataHub catalog. We obtained a set D of 293 triplesets whose VoID descriptions indicated the vocabularies, classes and properties the tripleset used. Out of the 42,778 possible links, we uncovered a set L of 410 links connecting such triplesets by analyzing the `void:linkset` property.

Ground truth. Due to the lack of benchmarks for validating the creation of links between triplesets, we adopted as ground truth the set L of links defined above. Furthermore, we separated the tripleset pairs in $D \times D$ into two classes: (i) *(ground truth) linked tripleset pairs* that are connected by a link in L, and (ii) *(ground truth) unlinked tripleset pairs* that are not connected by a link in L.

Performance measures. To validate the recommendation algorithms, we adopted the standard metrics **R**ecall, **P**recision and **F**-measure, defined based on true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) links between triplesets. Briefly, the *positive* and *negative* terms refer to



(a) (b) (c)

Fig. 1: (a) Precision, (b) Recall and (c) F-measure of the supervised classifiers by the percentage of (ground truth) unlinked tripleset pairs considered (100%, 75%, 50%, 25% and 1%)

link prediction, while *true* and *false* refer to the links in L. Thus, precision, recall and F-measure are defined as: $\mathbf{P} = \frac{TP}{TP+FP}$; $\mathbf{R} = \frac{TP}{TP+FN}$; and $\mathbf{F} = 2 \times \frac{\mathbf{P} \times \mathbf{R}}{\mathbf{P}+\mathbf{R}}$.

Baselines. As baselines for the experiments, we used two standard supervised learning algorithms: Support Vector Machines - SVM (LibLINEAR implementation) and Multilayer Perceptron. Similarly to the J48 decision tree, we used both SVM and Multilayer Perceptron to classify pairs of triplesets into (ground truth) linked tripleset pairs and (ground truth) unlinked tripleset pairs, based on link prediction measures values estimated considering different features sets.

Results. Before discussing the results, we observe that a pair of triplesets may not be in L, the set of links obtained from the DataHub catalog, because of a lack of metadata information or because they were never interlinked, but they might be. This indeterminacy might contaminate the learning algorithms. Hence, we vary the percentage of (ground truth) unlinked tripleset pairs considered when analyzing the performance of the various algorithms.

Figure 1 shows the precision, recall and F-measure achieved when the percentage of (ground truth) unlinked tripleset pairs varies (100%, 75%, 50%, 25% and 1%), while maintaining the number of (ground truth) linked tripleset pairs constant:

- Figure 1(a) shows that both the Multilayer Perceptron and the J48 implementations achieved a precision greater than 85%, independently of the percentage of (ground truth) unlinked tripleset pairs considered.
- Figure 1(b) indicates that the recall of the supervised classifiers increases when the percentage of (ground truth) unlinked tripleset pairs is reduced.
- Figure 1(c) shows that the J48 algorithm obtained the best overall performance, independently of the percentage of (ground truth) unlinked tripleset pairs considered.

To conclude, the J48 implementation achieved higher recall and F-measure, independently of the percentage of (ground truth) unlinked tripleset pairs considered.

4 Conclusions

In this paper, we presented a tool for tripleset recommendation, called *TRTML*, which reduces the effort of searching for related triplesets in large data repositories. *TRTML* is based on link prediction measures and supervised learning algorithms. The crucial role of the supervised learning algorithms is to automatically select a set of features, extracted from the VoID vocabulary, and a set of link prediction measures that, when combined, lead to effective tripleset interlinking recommendations. After a comprehensive evaluation of the supervised learning algorithms, the results show that the implementation based on the J48 decision tree (Quinlan's C4.5 implementation) achieved the best overall performance, when compared with the Multilayer Perceptron and the SVM algorithms.

Acknowledgments. This work was partly supported by CNPq, under grants 160326/2012-5, 303332-2013-1 and 557128/2009-9, by FAPERJ, under grants E-26/170028/2008 and E-26/103.070/2011.

References

1. Berners-Lee, T.: Linked Data - Design Issues (June 2009) W3C, http://www.w3.org/DesignIssues/ LinkedData.html, accessed on March 2013.
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. **5**(3) (Mar 2009) 1–22
3. Nunes, B.P., Kawase, R., Fetahu, B., Dietze, S., Casanova, M.A., Maynard, D.: Interlinking documents based on semantic graphs. In: KES. Volume 22 of Procedia Computer Science., Elsevier (2013) 231–240
4. Nunes, B.P., Mera, A., Casanova, M.A., Fetahu, B., Leme, L.A.P.P., Dietze, S.: Complex matching of rdf datatype properties. In: DEXA. Volume 8055 of LNCS. Springer (2013) 195–208
5. Nunes, B.P., Dietze, S., Casanova, M.A., Kawase, R., Fetahu, B., Nejdl, W.: Combining a co-occurrence-based and a semantic measure for entity linking. In: ESWC. Volume 7882 of LNCS., Springer (2013) 548–562
6. Fetahu, B., Dietze, S., Nunes, B.P., Casanova, M.A., Taibi, D., Nejdl, W.: A scalable approach for efficiently generating structured dataset topic profiles. In: ESWC, Springer *(to appear)* (2014)
7. Caraballo, A.A.M., Nunes, B.P., Lopes, G.R., Leme, L.A.P.P., Casanova, M.A., Dietze, S.: Trt-a tripleset recommendation tool. In: ISWC (Posters & Demos). (2013) 105–108
8. Lopes, G.R., Leme, L.A.P.P., Nunes, B.P., Casanova, M.A., Dietze, S.: Recommending tripleset interlinking through a social network approach. In: WISE. Springer (2013) 149–161