TRT - A Tripleset Recommendation Tool

Alexander Arturo Mera Caraballo[1], Bernardo Pereira Nunes[1], Giseli Rabello Lopes[1], Luiz André P. Paes Leme[2], Marco A. Casanova[1], Stefan Dietze[3]

[1] Department of Informatics, PUC-Rio, Rio de Janeiro/RJ – Brazil
{acaraballo,bnunes,grlopes,casanova}@inf.puc-rio.br
[2] Computer Science Institute, Fluminense Federal University, Niterói/RJ – Brazil
{lapaesleme}@ic.uff.br
[3] L3S Research Center, Leibniz University Hannover, Germany
{dietze}@l3s.de

Abstract. *According to the Linked Data principles, a tripleset should be interlinked with others to take advantage of existing knowledge. However, interlinking is a laborious task. Thus, users interlink their triplesets mostly with data hubs, such as DBpedia and Freebase, ignoring the more specific yet often even more promising triplesets. To alleviate this problem, this paper describes a tripleset interlinking recommendation tool based on link prediction techniques and evaluates the tool on a real-world tripleset repository.*

Key words: Linked Data, Recommender Systems, Social Networks

1 Introduction

A considerable number of triplesets, following the Linked Data principles, have already been published in a large number of areas, ranging from geographic to bibliographic data. This growth makes it difficult to choose which triplesets should be interlinked with a given tripleset. Thus, users interlink their triplesets mostly with data hubs, such as DBpedia and Freebase, ignoring the more specific triplesets which often contain particularly useful data. Furthermore, the metadata provided in data repositories such as the DataHub are typically not sufficient to help users choose the most suitable triplesets to interlink with.

To help alleviate this situation, we describe a tool for tripleset interlinking recommendation, based on previous work by the authors [1, 2]. More precisely, the tool addresses the *tripleset recommendation problem*, defined as follows: Given a tripleset t and a set of triplesets S, rank the triplesets in S based on the probability of interlinking t with them.

2 TRT - The Tripleset Recommendation Tool

Recommendation Procedure. A *tripleset* t is a set of RDF triples. A resource, identified by an RDF URI reference s, is *defined* in t iff s occurs as the subject of a triple in t.

Table 1: Local and quasi-local indices

Indice		**Equation**
Type	**Name**	
Local indices	Common Neighbors	$CN_{t,u} = \lvert C_t \cap C_u \rvert$
	Salton	$Salton_{t,u} = \frac{\lvert C_t \cap C_u \rvert}{\sqrt{C'_t . C'_u}}$
	Jaccard	$Jaccard_{t,u} = \frac{\lvert C_t \cap C_u \rvert}{\lvert C_t \cup C_u \rvert}$
	Sørensen	$Sørensen_{t,u} = \frac{2.\lvert C_t \cap C_u \rvert}{C'_t + C'_u}$
	Hub Promoted index	$HPI_{t,u} = \frac{\lvert C_t \cap C_u \rvert}{min\{C'_t, C'_u\}}$
	Hub Depressed index	$HDI_{t,u} = \frac{\lvert C_t \cap C_u \rvert}{max\{C'_t, C'_u\}}$
	Leicht-Holme-Newman	$LHN_{t,u} = \frac{\lvert C_t \cap C_u \rvert}{C'_t . C'_u}$
	Preferencial Attachment	$PA_{t,u} = C'_t . C'_u$
	Adamic-Adar	$AA_{t,u} = \sum_{w \in C_t \cap C_u} \frac{1}{log \lvert C'_w \rvert}$
	Resource Allocation	$RA_{t,u} = \sum_{w \in C_t \cap C_u} \frac{1}{\lvert C'_w \rvert}$
Quasi-local indices	Local Path	$LP_{t,u} = A^2 + \varepsilon A^3$
	Local Random Walk	$LRW_{t,u}(s) = \frac{C'_t}{2\lvert C \rvert} . \pi_{t,u}(s) + \frac{C'_u}{2\lvert C \rvert} . \pi_{u,t}(s)$

Let t and u be two triplesets. A *link* from t to u is a triple of the form (s, p, o), where s is an RDF URI reference identifying a resource defined in t and o is an RDF URI reference identifying a resource defined in u; we also say that (s, p, o), *interlinks* s and o. We say that t *can be interlinked* with u iff it is possible to define links from t to u. A *Linked Data network* is a graph $G = (S, C)$ such that S is a set of triplesets and C contains an edge (t, u), called a *connection* from t to u, iff there is at least one link from t to u.

Our recommendation procedure analyses the Linked Data network in much the same way as a Social Network. The inputs of the procedure are: (i) a *Linked Data network* $G = (S, C)$; (ii) a *target tripleset* t not in S (intuitively the user wishes to define links from t to the triplesets in S); and (iii) a *target context* C_t for t consisting of one or more triplesets u in S (intuitively the user knows that t can be interlinked with u). The output is an order list L of triplesets in S, called a *ranking*. The triplesets in the ranking are ordered using link prediction techniques discussed in what follows.

Link prediction techniques. The procedure uses link prediction theory to estimate the likelihood of the existence of a link between triplesets. We focus on local and quasi-local indices to measure the structural similarity between triplesets [3] according to their link structure. Table 1 summarizes the indices the procedure implements, where:

- C_i is the context of i (triplesets that i points to), where i a specific tripleset;
- C'_i is the inverse context of i (triplesets that point to i), where i a specific tripleset;
- A^j is the number of different paths with length j connecting t and u;
- ε is a free parameter;
- $\pi_{t,u}(s)$ is the probability that a random walker starting on t locates u after s steps;
- C is the set of all edges of the Linked Data network G.

Description of the TRT Tool in Action. Briefly, suppose that the user is working on a tripleset t and wants to discover one or more triplesets u such that t can be interlinked with u. He then uses the tool to obtain recommendations.

The tool first builds the Linked Data network $G = (S, C)$ defined by the metadata stored in the DataHub repository.

Then, the user defines the rest of the input data the tool requires. He may define a target context C_t for t, consisting of one or more triplesets in S, in two different ways: (i) by providing a VoID descriptor V_t for t from which the tool extracts C_t by analysing the *void:linkset* declarations occuring in V_t; or (ii) by manually selecting triplesets from the categories the tool displays. Finally, the user chooses a similarity index from those shown on Table 1.

From this input data, the tool outputs a ranked list of triplesets, thereby helping reduce the effort required to find related triplesets for the interlinking process.

The tool can be accessed at `http://web.ccead.puc-rio.br:8080/Uncover/`.

3 Evaluation

The tool was evaluated using the DataHub repository, which contains more than 6,000 triplesets, with approximately 15 thousand links that connect only 711 of the available triplesets. The links across triplesets were used to rank and recommend triplesets for interlinking. The recommendation process was assessed using the 10-fold cross validation approach, where we randomly divided the observed links into 10 subsets used as recommendation subgraphs. Finally, the overall performance was computed in terms of the average of the performances in the testing partitions.

To evaluate the prediction indices, we used three standard metrics: Area Under the receiver operating characteristic Curve (AUC), Mean Average Precision (MAP) and Recall. Table 2 summarizes the results for different target context sizes (shown in the first column of the table). The entries corresponding to the highest results among the 12 indices are emphasized in boldface underlined. The reader may observe that the PA index achieved the highest AUC (ranging from 83.74% to 95.90% depending on the target context size). The PA index also obtained the best MAP (37.83%) for target contexts with very few triplesets, while the RA index turned out to be more precise (72.42%) for larger target contexts. Table 2 also shows the coverage results. The PA index obtained the highest recall (96.4%), regardless of the size of the target context.

Table 2: AUC, MAP and Recall of the local and quasi-local indices

AUC	CN	Salton	Jaccard	Sørensen	HPI	HDI	LHN	PA	AA	RA	LP	LRW
1	70.52	47.79	69.84	69.28	48.94	69.31	48.00	**83.74**	71.31	70.53	70.74	69.67
5	87.10	55.73	81.20	80.93	58.78	80.17	52.24	90.76	88.45	88.02	**92.70**	83.21
10	92.42	57.14	85.06	84.85	60.84	83.79	52.87	**92.81**	92.37	92.40	92.25	86.69
20	92.77	58.47	88.34	88.30	59.45	86.54	51.39	**94.33**	92.53	92.64	92.76	88.22
50	92.84	59.10	92.96	92.99	56.27	92.09	52.30	**95.90**	92.17	92.72	91.91	90.26
MAP	CN	Salton	Jaccard	Sørensen	HPI	HDI	LHN	PA	AA	RA	LP	LRW
1	18.17	14.49	16.30	14.73	17.08	15.00	14.80	**37.83**	18.06	17.80	18.46	15.57
5	49.48	25.07	21.80	20.36	35.14	19.20	18.38	48.26	**52.20**	51.48	58.23	26.05
10	63.49	30.99	30.40	28.71	41.81	24.41	19.44	52.62	63.43	**63.71**	62.63	31.91
20	71.20	34.22	44.37	43.56	38.14	34.14	17.90	53.97	71.46	**72.38**	70.59	34.66
50	71.13	27.73	69.49	70.55	20.64	66.14	15.92	47.30	70.99	**72.42**	67.51	39.03
Recall	CN	Salton	Jaccard	Sørensen	HPI	HDI	LHN	PA	AA	RA	LP	LRW
1	48.72	49.69	49.86	49.76	49.68	49.55	50.02	**96.40**	50.81	48.74	48.81	49.12
5	81.45	83.80	82.69	83.03	83.68	82.23	82.43	**98.45**	83.63	82.83	86.90	82.42
10	89.52	88.73	89.42	89.29	89.35	89.85	89.17	**98.74**	89.49	89.28	88.96	89.21
20	90.03	90.31	89.68	89.18	89.12	89.53	89.19	**99.80**	89.50	89.58	89.84	90.01
50	90.05	90.16	90.15	90.21	90.04	89.38	88.45	**99.56**	89.06	89.58	89.02	89.71

4 Conclusions

In this paper, we proposed the use of link prediction techniques to address the tripleset recommendation problem in the Linked Data domain and presented a tool that implements the techniques. The tool computes local and quasi-local indices to predict links between triplesets. The results showed that the tool performs better, with respect to both AUC and recall, when the PA index is adopted. In terms of MAP, the PA index should be adopted for smaller context sizes, while the RA index should be adopted for larger context sizes.

Acknowledgments. This work was partly supported by CNPq, under grants 160326/2012-5, 301497/2006-0, 475717/2011-2 and 57128/2009-9, by FAPERJ, under grants E-26/170028/2008 and E-26/103.070/2011.

References

1. Leme, L.A.P.P., Lopes, G.R., Nunes, B.P., Casanova, M.A., Dietze, S.: Identifying candidate datasets for data interlinking. In Daniel, F., Dolog, P., Li, Q., eds.: ICWE. Volume 7977 of Lecture Notes in Computer Science., Springer (2013) 354–366
2. 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: Proceedings of WISE'13. (2013 (to appear))
3. Lü, L., Jin, C.H., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. Physical Review E **80**(4) (2009) 046122