What's all the Data about? - Creating Structured Profiles of Linked Data on the Web

Besnik Fetahu‡, Stefan Dietze‡, Bernardo Pereira Nunes*, Marco Antonio Casanova*, Davide Taibi†, and Wolfgang Nejdl‡

{fetahu,dietze,nejdl}@L3S.de, {bnunes,casanova}@inf.puc-rio.br, davide.taibi@itd.cnr.it

‡L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
*Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil
†Institute for Educational Technologies, CNR, Palermo Italy

Keywords

Profiling; Metadata; Vocabulary of Links; Linked Data

1. INTRODUCTION

The emergence of the Web of Data, in particular Linked Open Data (LOD) [1], has led to an abundance of data available on the Web. Data is shared as part of datasets, often containing inter-dataset links [6], mostly concentrated on established datasets, such as DBpedia[1]. Datasets vary significantly with respect to represented resource types, currentness, coverage of topics and domains, size, used languages, coherence, accessibility [3] or general quality aspects. The challenges from such diversity are underlined by the limited reuse of datasets from the LOD Cloud[2], where reuse and linking often focus on well-known datasets like DBpedia. Therefore, descriptive and reliable metadata are paramount to enable targeted search, assessment and reuse of datasets.

To address these issues and building up on earlier work [4], we propose an automated approach for creating structured profiles describing the topic coverage of individual datasets. The proposed approach considers a combination of sampling, topic extraction and topic ranking techniques. The sampling process is used to determine the best trade-off between scalability and profiling accuracy. Topic ranking is based on an adoption of graphical models *PageRank*, *K-Step Markov*, and *HITS*, which introduces prior knowledge into the computation of vertex importance [7]. Finally, the generated profiles are exposed as part of a public dataset based on the *Vocabulary of Interlinked Datasets* (VoID[3]) and the newly introduced vocabulary of links (VoL)[4] which describes the degree of relatedness between datasets and topics.

[1] `http://dbpedia.org`
[2] `http://datahub.io/group/lodcloud`
[3] `http://vocab.deri.ie/void`
[4] `http://data.linkededucation.org/vol/`

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577334.

2. PROFILING OF LINKED DATASETS

We propose a profiling pipeline which considers several steps for generating structured dataset profiles. The main steps are: (i) dataset metadata extraction from DataHub, (ii) resource type and instances extraction, (iii) entity and topic extraction, and (iv) topic ranking.

Step (i) makes use of the DataHub API, while in Step (ii) we investigated sampling strategies for extracting specific resource instances, as following: **'random sampling'** selects randomly resource instances from a dataset for further analysis in Steps (iii)-(iv); **'weighted sampling'** weighs each resource instance through the ratio of the number of datatype properties that define a resource over the maximum number of properties from all resources from a specific dataset; and **'centrality sampling'** weighs each resource through the ratio of the number of resource types used to describe it over the total number of resource types in a dataset. Similarly as for **'weighted sampling'** the weights determine the inclusion probability of a resource in the sample.

Step (iii) extracts entities by analysing the textual content of a resource using DBpedia Spotlight[5]. From the resulting entities, topics are extracted (as DBpedia categories) from the datatype property `dcterms:subject`.

In Step (iv), extracted topics are filtered out if the topic score is below the average score of all topics for a dataset, based on the *normalised topic relevance* score (see Equation 1), a variant of *tf-idf*. Topic ranking is based on *PageRank, K-Step Markov* (KStepM) and *HITS* scores, relying on an adoption from White and Smyth [7] which introduces prior knowledge in the computation of the mentioned algorithms, in our case representing datasets and their analysed resource instances. The models with priors are indicated as *PageRank with Priors* (PRankP), *HITS with Priors* (HITSP) accordingly. The resulting ranking based on the above models is adopted such that for each computed vertex importance in our dataset-topic graph, the weights are transferred into edge weights between topics and datasets (ranking topics for each dataset, with each dataset considered as prior knowledge).

$$NTR(t, D) = \frac{\Phi(\cdot, D)}{\Phi(t, D)} + \frac{\Phi(\cdot,\cdot)}{\Phi(t,\cdot)}, \quad \forall t \in \mathcal{T} \wedge D \in \mathcal{D} \qquad (1)$$

where $\Phi(t, D)$ is the number of entities assigned to topic t and dataset D, while $\Phi(t,\cdot)$ is the number of entities for all

[5] `http://spotlight.dbpedia.org/`

datasets $\mathcal{D}$ for t. $\Phi(\cdot, D)$ number of entities for D, and the number of entities $\Phi(\cdot,\cdot)$ for all $\mathcal{D}$.

3. EXPERIMENTAL SETUP

Data and Ground Truth: In our experiments we generated structured profiles for all LOD Cloud datasets where the respective endpoint was available. A subset of datasets $\mathcal{D}$={'lak-dataset', 'semantic-web-dog-food', 'socialsemweb-thesaurus', 'yovisto', 'clean-energy-data-reegle', 'oxpoints', 'courts-thesaurus'} are used as our *ground truth*. The corresponding dataset profiles consist of topics (DBpedia categories), extracted from resources that provide topic indicators in the form of *keywords* or *tags*. These term-based topic indicators were linked to DBpedia categories by extracting entities manually from the respective terms. The resulting topics were ranked according to their frequency $\forall D \in \mathcal{D}$.

Baselines: The baselines generate ranked topic profiles based on (i) *tf-idf* term weighting and (ii) *LDA* topic modelling [2]. To generate profiles consisting of categories from the sets of terms generated by the baselines, we extract entities and further link to categories such terms. The respective baseline term scores are used to rank the topics. For the baselines we analyse the full set of resource instances.

4. RESULTS AND EVALUATION

In this section, we compare the *profiling accuracy* from the topic ranking approaches in the profiling pipeline and those of the baselines against the *ground truth* profiles, measured using the `NDCG` metric. For *scalability* we analyse the trade-off between sample size and profiling accuracy. A demo of generated profiles is accessible at: `http://data-observatory.org/lod-profiles/profile-explorer`.

The *profiling accuracy* for the topic rankings is shown in Figure 1, considering all resource instances. Hence, the results from the various sampling strategies used in our profiling pipeline are equal. The `NDCG` scores are averaged over all datasets $\mathcal{D}$ and ranks. *PRankP* and *HITSP* were initialised with 10 iterations and parameter $\beta = 0.5$ (the probability of jumping back to a known vertex). For *KStepM* the number of steps was set to $K = 5$. Furthermore, the profiling accuracy results for the the topic ranking approaches *PRankP, HITSP* and *KStepM* correspond to the accuracy results when combined with the *NTR* score. The best performing approaches are *PRankP* and *KStepM* in combination with **'centrality sampling'**, having negligible difference. For the first baseline *tf-idf* the profiles were generated using the top-200 terms, while for *LDA* we used Mallet [5] with several initialisations, the best results were achieved using 20 topics and top-100 ranked terms.

In terms of *scalability* Figure 2 displays the trade-off between the amount of time taken to rank topics (excluding time taken to perform Step (ii)) and the analysed sample size. In details, the leftmost *y-axis* shows the log-scale of the topic ranking time for different sample sizes, while the rightmost *y-axis* shows the corresponding profiling accuracy. The best trade-off between *accuracy* and *scalability* is found at 5% and 10% sample sizes.

Acknowledgements: This work was partly funded by the LinkedUp (GA No:317620) and DURAARK (GA No:600908) projects under the FP7 programme of the European Commission.

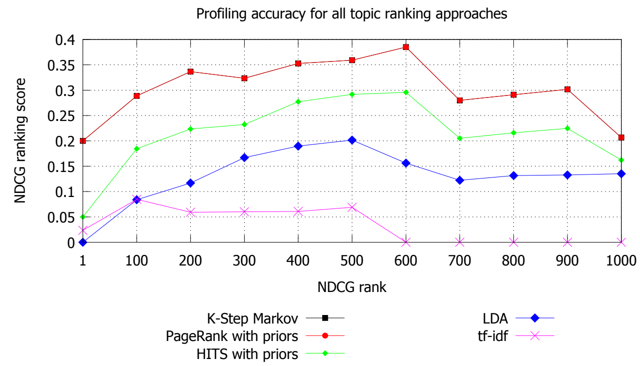

Figure 1: Profiling accuracy for all resources and `NDCG` averaged over all datasets in the *ground-truth*.

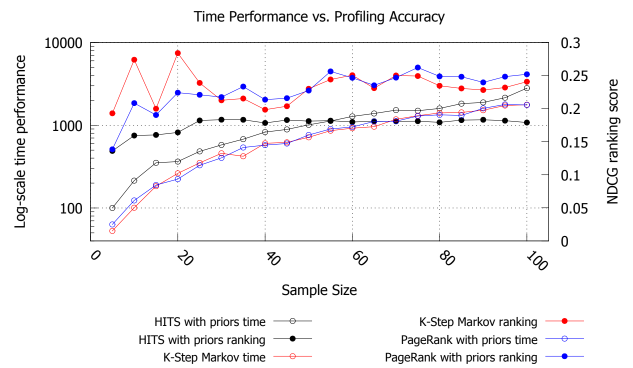

Figure 2: Trade-off between ranking time (in seconds) and profiling accuracy (with 'centrality sampling').

5. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *Proceedings of the 12th International Semantic Web Conference, Sydney, Australia*, 2013.

[4] B. Fetahu, S. Dietze, B. P. Nunes, D. Taibi, and M. A. Casanova. Generating structured profiles of linked data graphs. In *International Semantic Web Conference (Posters & Demos)*, pages 113–116, 2013.

[5] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[6] B. P. Nunes, S. Dietze, M. A. Casanova, R. Kawase, B. Fetahu, and W. Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC)*, pages 548–562, 2013.

[7] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the SIGKDD, 9th International Conference on Knowledge Discovery and Data Mining*, pages 266–275, 2003.