Towards Semantic Dataset Profiling

Mohamed Ben Ellefi, Zohra Bellahsene, François Scharffe, Konstantin Todorov
{firstname.lastname@lirmm.fr}

LIRMM / University of Montpellier 2, France

Abstract. The web of data is growing constantly, both in terms of size and impact. A potential data publisher needs to dispose with recapitulative information on the datasets available on the web, so that she can easily identify where to look for the resources to which her data relates. This information will help discover candidate datasets for interlinking. In that context, we investigate the problem of dataset profiling. We define a dataset profile as a set of characteristics, both semantic and statistical, that allow to describe in the best possible way a dataset by taking into account the multiplicity of domains and vocabularies on the web of data.
Keywords: Linked data, Dataset discovery, Dataset profiling.

1 Introduction

"With linked data, when you have some, you can find other, related data". This is a simplified view of the fourth principle of linked data[1]. Finding resources that can be reused or linked requires a framework for comparison between datasets. We propose to identify a dataset by its profile – a set of characteristics that allow to describe this dataset in the best possible way. In our understanding, a dataset profile should be based on three main criteria. (1) It should combine a versatile set of features that describe a dataset, (2) it should be generated in domain independent manner, i.e., the generation procedure should be applicable to any dataset from any domain, and (3) It should be generated automatically. A dataset usually relies on multiple different vocabularies to describe resources. A profile should reflect these vocabularies allowing for the application of ontology matching techniques in the dataset discovery and interlinking task.

There has been relatively little research dedicated to this task. Fetahu *et al.* [1] propose an approach for creating structured dataset profiles, where a profile describes the topic coverage of a particular dataset. In the topic extraction, the full textual content of a resource is analyzed from all its literals. Then, DBpedia Spotlight[2] is used as a named entity recognition and disambiguation tool.

Böhm *et al.* [2] introduce the notion of *k-similarity*, where two resources are *k-similar*, if k of their property/value combinations are exact matches. The intuition is that two resources are similar to some degree if they share a common set of attributes, and could therefore be related. The *k-similarity* approach can be seen as similar to a dataset profiling technique based on k property/value combinations.

[1] http://www.w3.org/DesignIssues/LinkedData
[2] http://spotlight.dbpedia.org

Atencia *et al.* [3] introduce a method for analyzing datasets based on key dependencies. This approach is inspired by the notion of a key in relational databases. A key in a dataset is a set of property/value pairs indicating that any resource in this dataset will have a unique set of values for a given set of properties. This definition of a key tolerates a few instances having the same values for the properties, what the authors have named a "pseudo-key" – a relaxed version of a key on the basis of a discriminability threshold. The pseudo-keys can be used to select sets of properties/values with which to compare resources issued from different datasets. Such a set of properties/values can be seen as a set of dataset features forming a profile.

While Fetahu *et al.* propose an automatic and domain independent profiling technique based on topics, the extracted profile is not intended to the comparison task. The other methods do not define the dataset profiling problem explicitly. Böhm *et al.* adopt an automatic selection of properties and the dataset profile is fixed after the comparison task. In particular, cross-vocabulary properties mapping is not performed which limits the space of semantic comparison. Atencia *et al.* appears to have less of the flaws of the other two methods, the set of properties being selected automatically and by taking into account the cross-domain context. However, the problem of cross-vocabularies pseudo-keys comparison is not discussed. Consequently, none of these techniques satisfies all the criteria and none of them addresses the main cross-vocabularies comparison problem. In next section, we give our proposition for a dataset profile based on these comparison criteria.

2 A Generic Framework for Dataset Profiling

A variety of characteristics can be included in a definition of a profile, such as a set of property/values pairs, types, topics or statistics. Several questions arise: What are the most representative features for a dataset? Are these features sufficient with respect to the dataset comparison task and where to look for them? We suggest that there are two main types of information that are relevant when constructing a dataset profile.

The first type is based on the declarative description of statistical dataset characteristics, such as property usage, vocabulary usage, datatypes used and average length of string literals, size, coverage, discriminability, frequency, etc. Thus, a framework like the LODStats [4] can be used as a large-scale dataset analytics which allows this kind of dataset profiling.

The second type is based on a set of types (schema concepts) that represent the topic and the covered domain. A set of properties/values describes the semantics of a dataset and enables the semantic comparison on an instance level.

The first type is appropriate to obtain information with regard to the structure, coverage and coherence of data. These statistical characteristics can be helpful to evaluate the dynamicity of a dataset in order to optimise reuse, revision or query tasks. In the current study, we are interested in characteristics of the second type only, since they provide explicit semantic information, useful for the dataset discovery task.

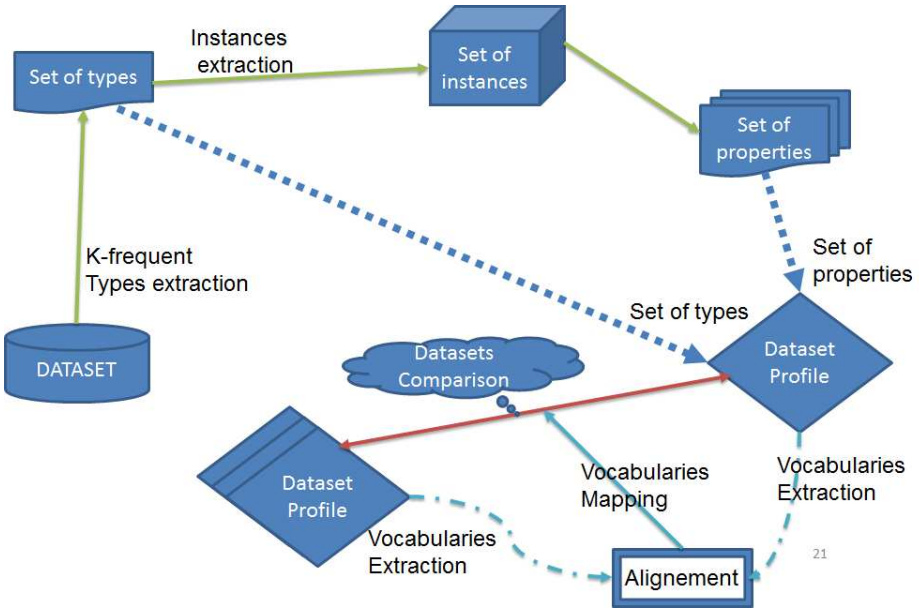

Fig. 1. Dataset Profiling Workflow

Our proposal is described in Figure 1. First, we select the *k-most* frequent concepts (types). We are inspired by [2] where the dataset content is summarized by the top ten discovered types. Then, we extract all instances associated to the selected set of concepts. Thereafter, the pseudo-keys generation technique [3] is applied to these instances, resulting in a set of property/values pairs. Thus, we can surpass the complexity problem of the pseudo-keys and we take advantage of the automaticity and the cross-domain features of the approach, assuring that the profile fulfills criteria 2 and 4. Note that for selecting a set of properties/values pairs, we can apply statistical measures like the property entropy [5], discriminability and frequency, thus including statistical information of the first type in the profiling process.

Finally, we address the problem of integrating cross-vocabulary datasets based on their profiles. The generated profiles are described by cross-vocabulary features. Basca et al. [6] defines a vocabulary as a simple "lightweight" ontology. Hence, in order to be able to compare profiles and measure similarities, a system needs to "know" the correspondences between the types of resources. For example, we consider two datasets, one described using *FOAF*, the other using *VCard*. When comparing resources of these types, the properties *foaf:givenname* should be compared to *vcard:fn*, as well as the property *foaf:familyname* - to the property *vcard:ln*. The proposition is to establish correspondences between the different features described by different vocabularies allowing for more precise semantic comparison and, consequently – for a more representative seman-

tic dataset profile. D'Aquin et al. [7] propose the manual alignments between schemas as that seems less costly than an automatic alignment and they consider only the most frequent types and properties. Our tendency is converging towards the automatic alignment for only popular vocabularies, e.g., http://schema.org/.

3 General Discussion

Due to the inherent heterogeneity of linked data, an efficient profiling technique is necessary in the context of dataset comparison and candidate dataset identification for the interlinking process. Our study of the existing techniques has helped identify several problems that have to be addressed in future. Most importantly, the extraction of profile features has to be automated as much as possible and made applicable for all domains and all vocabularies. In order to deal with a dataset comparison based on the semantic content, the profile features have to be semantically descriptive. The very question of the actual set of features that best describe a dataset is a subject of ongoing analysis.

Our proposal potentially meets all criteria listed in the introductory section. We repose on both semantic and statistical features and show their connexion. The process of semantic features generation is totally independent of the dataset domain and can be performed automatically on the basis of discriminability, frequency and support. Finally, the use of ontology matching techniques, ensures the compatibility of profiles defined for cross-vocabulary datasets comparison.

In future, we plan to implement our proposal and test it on existing datasets on the web of data and in the context of the Datalyse project[3].

References

1. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl, "A scalable approach for efficiently generating structured dataset topic profiles," in *In Proceedings of the 11th Extended Semantic Web Conference*, Springer, 2014.
2. C. Böhm, J. Lorey, and F. Naumann, "Creating void descriptions for web-scale data," *J. Web Sem.*, vol. 9, no. 3, pp. 339–345, 2011.
3. M. Atencia, J. David, and F. Scharffe, "Keys and pseudo-keys detection for web datasets cleansing and interlinking," in *EKAW*, pp. 144–153, Springer, 2012.
4. S. Auer, J. Demter, M. Martin, and J. Lehmann, "Lodstats – an extensible framework for high-performance dataset analytics," *EKAW*, p. 353, 2012.
5. T. Gottron, M. Knauf, S. Scheglmann, and A. Scherp, "A systematic investigation of explicit and implicit schema information on the linked open data cloud," *The Semantic Web: Semantics and Big Data*, pp. 228–242, 2013.
6. C. Basca, S. Corlosquet, R. Cyganiak, S. Fernndez, and T. Schandl, "Neologism: Easy vocabulary publishing," in *SFSW*, pp. CEUR–WS.org/Vol–368, 2008.
7. M. D'Aquin, A. Adamou, and S. Dietze, "Assessing the educational linked data landscape," in *(WebSci), Paris, France*, pp. 43–46, ACM, 2013.

[3] This research is funded under the Datalyse project (http://www.datalyse.fr/)