Position Paper: Dataset profiling for un-Linked Data

Emilia Kacprzak[1,2], Laura Koesten[1,2], Tom Heath[1], and Jeni Tennison[1]

[1] The Open Data Institute, London, United Kingdom
{emilia.kacprzak,laura.koesten,tom.heath,jeni.tennison}@theodi.org,
[2] University of Southampton, Southampton, United Kingdom

Abstract. The vast amount of data on the web presents a growing need to advance data search. Rich and meaningful metadata can enhance the discovery of datasets and establish connections between them. Where metadata is not comprehensive, it can be expanded through dataset profiling. The relative importance of different types of profiles varies depending on the user's context and the objective of the task. We discuss an approach to find un-Linked datasets and increase result relevance by offering related information. We propose generating rich profiles for datasets; counting the number and strength of relations between them and showing a graph of profiles that represents connections between different datasets. We can thereby capture correlations between datasets that can then improve the efficiency and effectiveness of data search. If developed further this would improve discoverability and reusability of datasets.

Keywords: Profiles, Metadata, Data search, Discoverability, Dataset

1 Introduction

The availability of data is growing rapidly and with it, big data technologies and services. The vast amount of information available presents unique challenges around data integration, data ownership or assuring data quality to users [1]. One of the key challenges is the hidden character of many published datasets - this can be ascribed to a number of factors, one of which is the lack of useful metadata. Data search techniques are strongly dependent on adequate information about the datasets [2]. Searching is based on meaningfully indexed content, hence profiles are needed that describe the data in a manner that is most useful for discoverability.

Value derived from data is generated to a great extent through understanding relations between different datasets and between the entities they describe. Most techniques for extracting information from datasets rely on the statistical characteristics of the data or do not scale well [3]. Improving the quality and comprehensiveness of profiles can enable the discovery of datasets as well as enable the development of links between them; be they Linked or un-Linked data.

The remainder of this paper is structured as follows. Section 2 describes related work that influenced the concept presented in section 4. Section 3 gives a rationale for the focus on un-Linked data and presents use scenarios.

2 Related work

2.1 Profiling techniques

Data profiling comprises a broad range of methods to efficiently describe datasets [3]. Analysis of individual cells or columns is used to create summaries of data [4]. Relationships between values in the data are described as association rules; which have been suggested as additional profiling activities.

Naumann proposes joining two datasets based on their profiles, similar to a reverse engineering process which reveals possible relations between datasets based on constraints identified in the data [3]. Cell-level and column-level analysis enables effective schema matching between datasets. Thereby initially not related schemas can be mapped and semantically correct correspondences can be revealed [5]. Integration of seemingly unconnected datasets can provide additional insights on the original data.

These profiling techniques are strongly focused on numerical data and creating associations between datasets columns. On the other side of the spectrum, Linked Data is used to improve profiling processes. Fetahu et al. propose building a graph of linked datasets including the relative importance of each node, each of which is an openly available dataset. This is done with the objective of improving profiling accuracy [6]. Therefore, they developed an algorithm which adapts PageRank, K-Step Markov, and HITS models along with topical profiles generated with DBpedia. Profiles are enriched by the use of Linked Data (ibid.). Our approach is complementary to this, attempting to create links between datasets that are not already linked, by use of their profiles; as can be seen in section 4.

Users should be involved in the process of searching for data and richer profiles can support that search process. Elefi et al. emphasise the usefulness of explicit semantic information for the dataset discovery task to support the interlinking of datasets [7]. Wagner et al. present an approach to improve Linked Data search by entity-based data-source contextualization. In this they propose involving the user in the dataset selection process and place emphasis on the importance of not only offering exactly matching sources, but offering supplementary information from additional sources to provide context. Additional sources are defined as being relevant to the user's information need, rather than to the exact query. The proposed framework in section 4. follows a similar objective, but is not limited to Linked Data.

2.2 Types of profiles

The most basic type of profiles are structural profiles for which no external resource analysis is needed. These are created by reporting the number of columns,

rows and their data types, along with information about uniqueness and completeness of the data [4, 3]. This leads to explicit single value analysis which is then often defined as a numerical or statistical profile. To generate additional insights, topical profiles provide connections to different resources, for example DBpedia, by evaluating the datasets topic coverage [6]. Spatial profiles describe geographical data [8]. If the geographic position could be expressed as a place it can be used in the same way as topical profiles.

Although these different types of profiles exist, they were created to solve particular problems and are rarely used together. If combined meaningfully, these profiles could enhance discoverability of datasets even further. Section 4 describes how the availability of richer profiles could be used to understand connections between datasets.

3 Finding connections in un-linked data

3.1 Why un-linked data?

Existing research on techniques for understanding the relationships between datasets, such as [2],[9],[10] are focused on Linked Data. However, most data that can be found on data portals is not published as Linked Data. An analysis of open data available on data.gov.uk shows that out of all that are published under an Open Government license, less than 2% are RDF (12,470 open datasets published of which 213 are in RDF, 1.7%). This number excludes non-machine-interpretable formats such as PDF and HTML [11]. Hence, we argue that methods contributing to the improvement of profiles for all data formats enable discoverability and ultimately linking of data.

3.2 Use scenarios

We are focused on three use scenarios that address different user needs. These serve as high-level exemplification of the variety of issues users can encounter when attempting to find datasets. Apart from the type of data, we note that the context and the individual objective when searching for datasets influences the relevance of a specific type of profile for a specific task.

Scenario A Posing a question that can be answered by consulting a single data set. This scenario implies limited time and there should be no, or very limited, cost associated. Accuracy and licensing can be of varying importance. An example would be looking for the closest free wifi to the users current location. The profile should give information about whether or not this dataset contains data about my location and how up to date it is.

Scenario B Looking for datasets to compare for relevance and quality to be integrated into an application. Time and cost can vary in importance. Here

information about granularity and reusability (licensing, format) of the data can be more important. For example, if a user wants to evaluate the air quality in London in 1995 in comparison to today and the metadata mentions the years 2000 to 2005, a naive search for the year 2004 might not provide results.

Scenario C Comparison of as many as possible available datasets on a topic for research purposes. Here, more time is available and cost is less of an issue. Archived data can be important; coverage and granularity are selection criteria. For example, a user may want to compare all datasets available on a specific city to evaluate which areas to support financially.

What is evident from these scenarios is that considering the context of the user and the objective of the task, search process and source selection may vary. Furthermore the presentation of results and the way different users might want to interact with those results is also dependent on these factors. The proposed approach in section 4 serves as an initial step to resolve the issues described in the above scenarios, by allowing user interaction in the process of source selection.

4 Connecting un-linked data

As we have discussed, structural and statistical profiles can enhance the discovery of relations between datasets. Knowing the general structure of the dataset with its data types and value ranges can be the first step in narrowing down the number of columns that can relate to each other. This can be used to create actual links between unlinked datasets, e.g. by creating foreign keys between them [4].

Other types of profiles can be used for linking datasets in a broader, less strict, sense. Topical and spatial summaries are descriptions of datasets on a semantic level. Similar topics, coverage or data patterns can indicate relationships between datasets that could inform search.

We propose to build a weighted graph based on the profiles of un-Linked data. By generating rich profiles for datasets and counting the number and strength of relations between them,correlations between datasets could be captured which can then improve the efficiency and effectiveness of search within open data portals, such as data.gov.uk.

In accordance with Wagner et al. [9], we propose offering a broad range of results including contextual information, as well as the possibility of user involvement during source selection. The search process would be based on number and strength of connections between the profiles; after an initial search, the weighting of these individual connections could be influenced by the user.

In the first step, based on a query, keyword extraction can be used to generate ranked keywords which, as well as informing an initial search, are presented to the user in order of importance. The user can interact with the list of keywords, by swapping them with each other, to achieve more relevant results.

The relevance of the keywords to the user influences the subset of datasets that appear within this graph - not all keywords are equally important to the user. Results that match with all keywords are of higher importance; datasets that match with only a few would be included if there were not enough results. Fig. 1(a) shows an example query about the population of London in 1995 and the datasets that matched the query the most. It also illustrates the strength of the connections between keywords and the profiles of the suggested datasets. This increases the end users awareness of the dataset content before opening it. As a second step, a graph of connections between related datasets is presented,

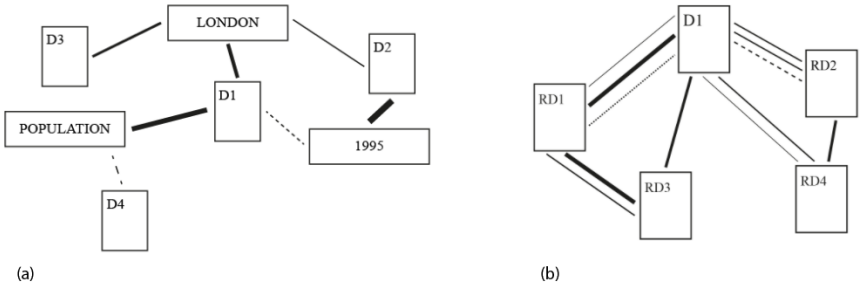

Fig. 1. (a) Graph of connections between keywords and datasets, based on topic relevance (b) Results presentation of graph suggesting related datasets

which is based on one chosen dataset from the results in fig 1(a). The weighting of connections between datasets are based on the similarity of the their profiles and the chosen datasets that matched the keywords from the original query, taking into account the users priorities. These initially identified datasets serve as recommendations and aid exploration of related information.

Fig. 1(b) presents results after the user chose the dataset D1 to be the most relevant for their information need. To improve the search process, other profiles which have strong connections to the profile of D1 are presented along with information about the type of the connections between them.

The graph should present an adjustable number of results in a graphical visualisation. The argument for the graph is to deepen the user's understanding and the mental model of the search process. The availability of relevant data depends on the topic as well as the specificity of the search. Hence the number, the availability and the similarity of the datasets presented as results can vary significantly. Data profiling is naturally a user-oriented task [3] and we argue that following on from that, the user should be involved in source selection to provide more advanced data search.

5 Conclusions and future work

Profile features vary in their importance dependent on the users context and objective. When looking at the use scenarios, section 3, different factors can be

important for users. Concerning the interaction with the system these concern time, cost, computational power, trust. Beyond that users have varying requirements on the type of information covered in metadata, depending on their task and context.

The approach described here would improve discoverability and aids all of the mentioned use scenarios as the user can be involved in the process of weighting the results. This can decrease the importance of the accuracy of ranking algorithms and enhance the quality of results according to ones individual needs. The possibility to detect similarities between profiles provides the basis for explorative search, engaging the user in source selection. To further develop this approach and validate its applicability we aim to develop profiles that support the identification of connections between datasets and topics, and between datasets and other datasets. Following from that we aim to develop interfaces, such as user-dependent result presentations, that support explorative data search. As the proposed graph is created on an abstract level and the connections are formed based on the datasets profiles the approach is not limited to a type of data, but can be used for Linked Data and un-Linked data at the same time; as the focus is to improve user experience.

References

1. Nasser, T., Tariq, R. S.: Big Data Challenges. J. Comput. Eng. Inf. Technol. vol. 4:3, 1000135 (2015)
2. Assaf, A., Troncy, R. , Senart, A.: Roomba: An extensible framework to validate and build dataset profiles. In: The Semantic Web: ESWC 2015 Satellite Events, pp. 325-339. Springer International Publishing (2015)
3. Naumann, F.: Data profiling revisited. ACM SIGMOD Record 42, no. 4., pp. 40-49. (2013)
4. Abedjan, Z., Golab, L., Naumann, F.: Profiling relational data: a survey. The VLDB Journal, vol 24(4), 125. (2015)
5. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10: 334350. (2001)
6. Fetahu B., Dietze S., Nunes B.P., Casanova M. A.,Taibi D, Nejd W.l.: A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles (2014)
7. Ellefi, M. B., Bellahsene Z., Scharffe F. ,Todorov K. Towards Semantic Dataset Profiling. In: PROFILES@ ESWC (2014)
8. Shekhar, S., Celik, M., George, B., Mohan, P., Levine, N., Wilson, R. E., Mohanty, P. Spatial analysis of crime report datasets. National Science Foundation (NSF), Washington DC (2010)
9. Wagner, A., Haase P., Rettinger A., Lamm, H.. Entity-based data source contextualization for searching the Web of data. In: The Semantic Web: ESWC 2014 Satellite Events, pp. 25-41. Springer International Publishing (2014)
10. Boehm, C., Lorey J., Naumann F.: Creating void descriptions for web-scale data. Web Semantics: Science, Services and Agents on the World Wide Web 9, no. 3, pp. 339-345. (2011)
11. Data.gov.uk., Beta: Datasets. 15/03/16: https://data.gov.uk.