Undefined 1 (2009) 1–5
IOS Press

DRX: A LOD browser and dataset interlinking recommendation tool

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Alexander Arturo Mera Caraballo, Bernardo Pereira Nunes and Marco A. Casanova
Department of Informatics, PUC-Rio
Rua Marquês de São Vicente, 225 Gávea, Rio de Janeiro, RJ, Brazil, Zip code: 22451-900
E-mail: {acaraballo,bnunes,casanova}@inf.puc-rio.br

Abstract. With the growth of the Linked Open Data (LOD) cloud, data publishers face a new challenge: finding related datasets to interlink with. To face this challenge, this paper describes a tool, called DRX, to assist data publishers in the process of dataset interlinking and browsing the LOD cloud. DRX is organized in five main modules responsible for: (i) collecting data from datasets on the LOD cloud; (ii) processing the data collected to create dataset profiles; (iii) grouping datasets using cluster algorithms; (iv) providing dataset recommendations; and (v) supporting browsing the LOD cloud. Experimental results show that DRX obtains good overall MAP when applied to real-word datasets, which demonstrates the ability of DRX to facilitate dataset interlinking.

Keywords: Dataset Recommendation, Dataset profiling, Dataset Clustering, Linked Data, Semantic Web

1. Contextualization

Despite the efforts to foster publishing data as Linked Open Data (LOD) [7], data publishers still face difficulties to integrate their data with other datasets available on the Web [3]. Defining RDF links between datasets helps improve data quality, allowing the exploration and consumption of the existing data.

Although frameworks that help discover RDF links are available, such as LIMES[1] [15] or SILK[2], the selection of the source and target datasets to be interlinked is still a manual, often non-trivial task. In what follows, we refer to this task as *dataset interlinking* and to the problem of suggesting a list of datasets to be interlinked with a given dataset as the *dataset interlinking recommendation problem*.

A review of the literature reveals relatively few studies dedicated to the dataset interlinking recommendation problem. For instance, Leme et al. [11] created a straightforward method based on the naïve Bayes classifier to generate a ranked list of related datasets. The relatedness between datasets was measured solely using *linksets*, a set of existing links between datasets, retrieved from the Datahub[3] catalog. Similarly, Lopes et al. [12] took advantage of *linksets* to provide dataset interlinking recommendations. They used link prediction measures, borrowed from the social network analysis field, to estimate the probability of datasets being interconnected.

Unlike these works, Nikolov et al. [14] investigated the use of a existing Semantic Web index (Sig.ma[4]) to identify candidate datasets for interlinking. Sig.ma is queried with text literals extracted from *rdfs:label*,

[1] http://aksw.org/Projects/LIMES.html
[2] http://silk-framework.com/
[3] http://datahub.io
[4] http://sig.ma/

foaf:name and *dc:title* properties from a given dataset to find the most overlapping datasets w.r.t to instances. Instead of using instances, Emaldi et al. [4] relied on the structural characteristics of datasets using a frequent subgraph mining (FSM) technique to identify and possibly establish links between disparate datasets. FSM is an interesting alternative to provide a more efficient approach as it only uses the most frequent subgraphs from a dataset to perform the analysis.

Dataset profiling/summarization techniques are also related to dataset interlinking recommendation. These techniques aim at elaborating a concise but comprehensive version of datasets. Thus, techniques such as those proposed by Lalithsena et al. [10] may ease the dataset interlinking process. They use reference datasets such as Freebase and DBpedia to enrich a set of instances from a given dataset to create a general description. A similar approach is proposed by Fetahu et al. [5], which created structured dataset profiles. Their approach combines several techniques, such as sampling, named entity recognition, topic extraction methods [13] and ranking [18], to represent a dataset. A more generic approach to create profiles on the Web is presented by Kawase et al. [9].Their approach aims at generating histograms for any text-based resource on the Web based on the 23 top-level categories of the Wikipedia ontology.

As previously stated, there is a limited number of studies devoted to the dataset recommendation problem, and this number decreases when it comes to tools. For example, TRT [1] provides a recommendation method that analyses the Linked Data network in much the same way as a Social Network [12]. That is, link prediction theory is used to estimate the likelihood of the existence of a link between datasets instead of users in a common Social Network scenario. TRTML [2] combines supervised learning algorithms and link prediction measures to provide dataset recommendations. The recommendation relies on the similarity between vocabularies, classes and properties used in disparate datasets.

In this paper, we present an overview of the DRX tool and illustrate its potential applications. DRX assists data publishers in the process of dataset interlinking and browsing the LOD cloud. DRX takes advantage of various methods including crawling, profiling, clustering and ranking modules to create ranked lists of datasets to be interlinked with a given dataset.

In more detail, when a data publisher wants to select datasets from the LOD cloud to set up a link discovery framework, only textual resources needs to be provided from the new dataset. The tool outputs a list of datasets, available in the LOD cloud, that most likely contain resources that can be interlinked with the resources of the new dataset. The tool also provides rich representations of the dataset profiles, which facilitate exploring the LOD cloud.

The tool incorporates the following main modules: (i) collecting data from datasets on the LOD; (ii) processing the data collected to create dataset profiles; (iii) grouping datasets using cluster algorithms; (iv) providing dataset recommendations; and (v) supporting dataset browsing. The results show that DRX has a good potential to be used as a dataset interlinking facilitator.

The remainder of this paper is organized as follows. Section 2 introduces the architecture of the tool and describes in details its five modules. Section 3 presents the DRX features through a case study. Section 4 describes the experiments conducted and discusses its outcomes. Finally, Section 5 concludes the work and presents suggestions for future work.

2. The DRX architecture

DRX is based on five modules, depicted in Figure 1, which are distributed in three different layers: data acquisition, data processing and application.

These modules perform five main tasks:

1. Collect data from datasets in the LOD cloud.
2. Process the data collected to create dataset profiles, called *fingerprint*.
3. Group the *fingerprint*, using cluster algorithms.
4. Provide dataset recommendations.
5. Support browsing the dataset profiles.

The data acquisition layer includes the *crawling* module, which discovers metadata about the LOD datasets from LOD catalogs, such as Datahub[5] and the Mannheim[6] catalog, as well as from manually submitted data. LOD catalogs typically stores metadata such as maintainer, SPARQL endpoint, relationships, VoID[7], tags, license and resources. The crawling module uses the CKAN[8] API to query metadata available in such catalogs.

[5] http://datahub.io

[6] http://linkeddatacatalog.dws.informatik.uni-mannheim.de

[7] http://www.w3.org/TR/void/

[8] http://ckan.org/

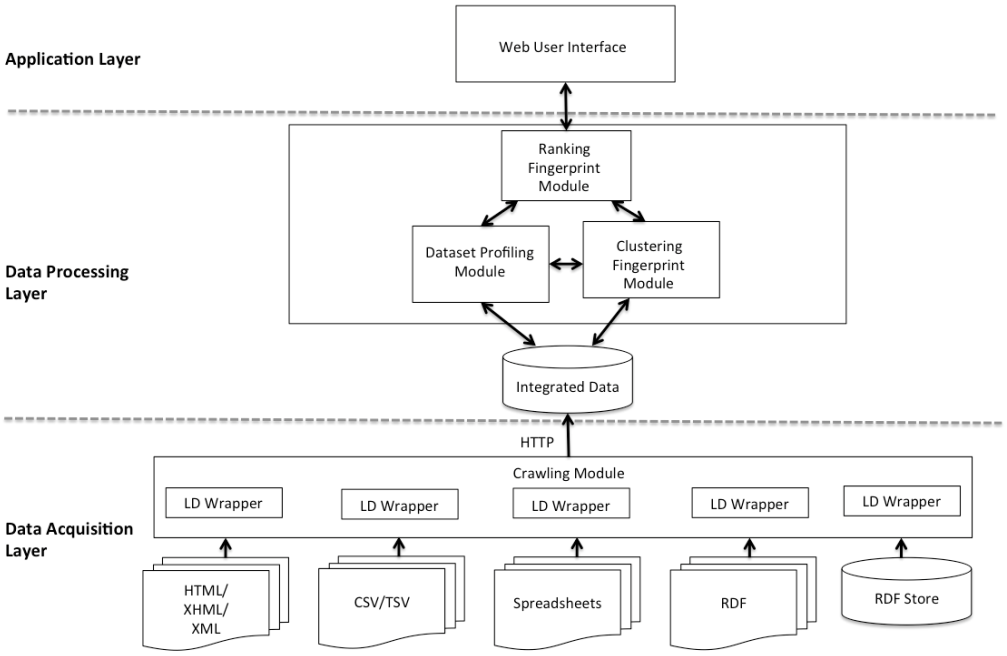

Fig. 1. Architecture DRX tool.

Since data can be in many different formats, the data acquisition layer also provides a set of specialized Linked Data wrappers, that the crawling module uses to extract textual resources from the datasets. Once a dataset is located, the crawling module creates a document containing the retrieved textual resources; the module ignores dataset with no textual resources, since the technique implemented in the profiling module, discussed in what follows, requires textual resources.

The data processing layer includes three main modules: *profiling*, *clustering* and *ranking*.

The profiling module processes the documents retrieved from a dataset and computes a description that characterizes the content stored in the dataset. DRX implements the technique described in [9], that generates dataset profiles or *fingerprints* from textual resources.

The technique has five steps:

1. Extract entities from a given textual resources.
2. Link the extracted entities to Wikipedia articles.
3. Extract categories for the articles.
4. Follow the path from each extracted category to its top-level category and compute a vector with scores for the top-level categories thus obtained (such as agriculture, applied science, arts, belief, business, chronology, culture and so on).
5. Perform a linear aggregation in all dimensions of the vectors to generate the final profile, represented as an histogram for the 23 top-level categories[9] of the English Wikipedia.

The clustering module groups together fingerprints that share a certain similarity. The top-level categories of the English Wikipedia act as a set of features. The DRX tool implements the XMeans cluster algorithm, which is part of the WEKA[10] (Waikato Environment for Knowledge Analysis) suite. The XMeans algorithm extends the K-means algorithm and includes an efficient estimation of the number of clusters [6,16].

Finally, the last module provides two strategies to provide recommendations for a given dataset d_t: *Cluster-based* and *Profiling-based* strategies. The main difference between both strategies is that the first strategy is limited to the datasets in a cluster whereas the profiling based strategy considers all datasets identified by fingerprints. Independently of the strategy chosen, for a given dataset d_t, the dataset recommendation module outputs a list of datasets ordered by the probability of being interlinked.

[9] https://en.wikipedia.org/wiki/Category:Main_topic_classifications

[10] http://www.cs.waikato.ac.nz/ml/weka/

Assume that d_t is in cluster C_{d_t} . The cluster-based strategy creates a ranked list by taking into account only the distance between the fingerprint of d_t and the fingerprints of the other datasets in C_{d_t} . By contrast, the profiling-based strategy creates a recommendation list based on the distance between the fingerprint of d_t and the fingerprints of all other profiled datasets. Additionally, the recommendation module is prepared to use other distance measures, such as Euclidean Distance, to create a ranked list of datasets.

3. DRX GUI and Case Study

To support the exploration and consumption of datasets in the LOD cloud, the DRX tool provides rich Web interfaces that help data publishers locate datasets relevant to their interests. DRX allows the analysis of the LOD cloud by using dendrograms, tables and coordinate graphs (see Figure 2). It does not require any expertise in Semantic Web technologies or languages.

To better illustrate the dataset recommendation approach, we selected an independent dataset, created by a reliable third party. The case study dataset, rkb-explorer-newcastle[11], was created jointly with other datasets under the ReSIST[12] project funded by European Union Work Programme. These datasets are available through the RKBExplorer[13] Semantic Web browser that supports the Computer Science research domain. It combines information from multiple heterogeneous sources, such as published RDF sources, personal Web pages, and data bases in order to provide an integrated view of this multidimensional space.

To start using the features that DRX offers, a user needs to load the Web application located in the following address `http://www.inf.puc-rio.br/~acaraballo/DRX`.

Users can be in two different scenarios, as follows:

- The user wants to register a new dataset and then browse the LOD to obtain recommendations regarding the new dataset;
- The user wants to browse the LOD to obtain recommendations for an existing LOD dataset.

The goal of the first step is to collect textual resources from data sources. Therefore, in the first scenario, the tool, through the register form feature (see Figure 2(a)), allows users to enter information such as the dataset name, dataset owner and textual resources, in order to register a new dataset.

In the second scenario, the crawling module has already collected textual resources, with the help of LD wrappers. In the case study, the dataset was crawled, using its SPARQL endpoint, to extract textual resources from the literal values of the **rdfs:label**, **skos:altLabel** or **skos:prefLabel** properties.

In what follows, the interaction between the user and the tool is the same, regardless of scenario. Thus, we will not distinguish them.

The second step is carried out transparently to the user. Here, textual resources collected in the first step are used as input to the profiling module to create a description of the content through a fingerprint.

Table 1 presents the fingerprint generated for the case study dataset, where the 23-dimension vector shows peaks for “Society”, “Technology” and “Science”, categories that are strongly related to the data content that rkb-explorer-newcastle dataset provides.

To facilitate the exploration and selection of datasets, it is important to reduce the search space of datasets in the LOD cloud. Therefore, the aim of the third step is to generate groups of datasets that share a certain similarity. The clustering module then implements a simple interface that allows users to enter input parameters (such as the minimum and maximum number of clusters and the number of seeds (see Figure 2(b)) and to execute the clustering process for all collected LOD datasets.

In the case study, we used a minimum of 8 clusters, since this is the number of categories of the LOD diagram[14]. The maximum number of clusters was set to 10 and the number of seeds was set to 10.

The results of the clustering process (see Figure 2(c)) are illustrated using a dendrogram representation that allows users to navigate among the generated clusters and their respective members. For the case study, the clustering process generated 10 groups and the rkb-explorer-newcastle dataset was clustered under cluster #3.

The user interface also offers a zoom-in/out feature which allows users to explore the members of each cluster in more detail; the user has to click inside a

[11] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/dataset/rkb-explorer-newcastle`

[12] `http://www.resist-noe.org/`

[13] `http://www.rkbexplorer.com/`

[14] `http://lod-cloud.net/`

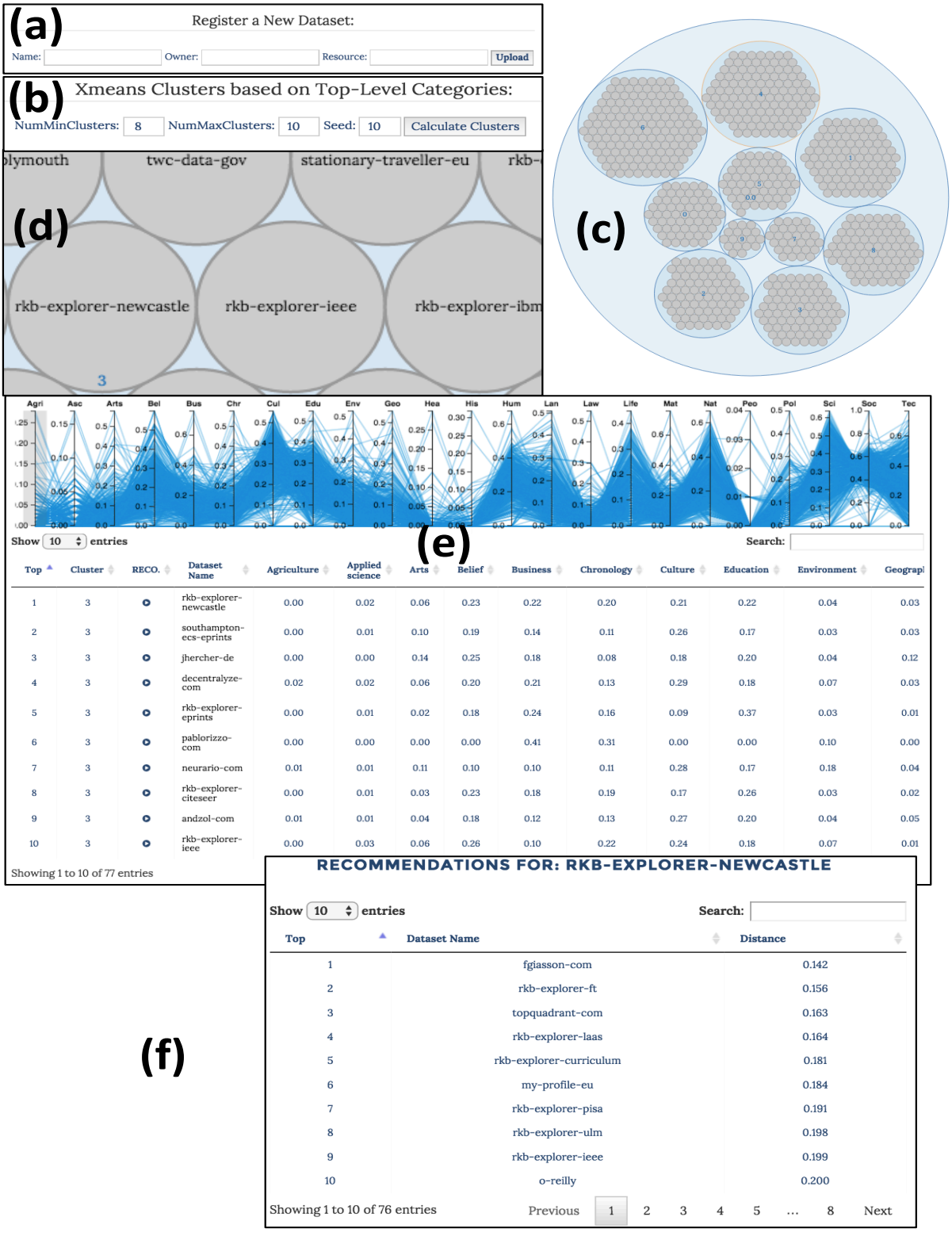

Fig. 2. Interface DRX tool.

cluster area to zoom-in the cluster(see Figure 2(d)). For example, after zooming in cluster #3, we may observe some of its member, such as rkb-explorer-ieee, twc-data-gov, aspire-plymouth and rkb-explorer- newcastle.

To provide an easy way to read and understand dataset profiles, the user interface provides a table with relevant information about the members of a selected cluster (see Figure 2(e)). This table offers column sorting, by selecting one of the columns, and full text search. Each row shows the following information: the *"top"* column provides a dataset ranking based on the centrality of the datasets in the cluster; the *"cluster"* column represents the cluster membership; the *"RECO."* column provides recommendations, for the dataset d_j on the selected row, from datasets of the same cluster as d_j; the *"dataset name"* column is a link to the dataset page in the Mannheim catalog; and, finally, the other columns show the vector with the 23 top-level categories.

For the case study, regarding cluster #3, Figure 2(e) shows detailed information of 10 members of cluster #3 out of a total of 77. Additionally, the rkb-explorer-newcastle dataset was assigned the first position in the list, since it has the highest centrality degree in the cluster.

Finally, the user interface offers a feature to obtain interlinking recommendations. The user simply selects the dataset from the table in Figure 2(e) and then click on the corresponding cell of column *"RECO."*. Then, a table is displayed, containing a list sorted by ascending order of the Euclidean distance values (see Figure 2(f)). For the case study, 10 recommendations, of a total of 76, are displayed (see Figure 2(f)). Note that the top ten recommendations include 6 datasets from the project that rkb-explorer-newcastle belongs to: rkb-explorer-ft, rkb- explorer-laas, rkb-explorer-curriculum, rkb-explorer- pisa, rkb-explorer-ulm and rkb-explorer-ieee.

4. Evaluation

4.1. Description of the Data and the Evaluation

The approach that DRX implements was assessed using data retrieved from the Mannheim catalog, a metadata repository for open datasets. Through the CKAN API, the catalog enables querying dataset metadata, including two multivalued properties (*relationships* and *extras*), which in turn allows data publishers to assert that a dataset points to another one. Both properties were used to retrieve the linksets between datasets in the Mannheim catalog. During the crawling step, we strictly retrieved datasets that had at least one defined linkset. In mid 2015, the data collected amounts to 510 datasets, with a total of 8,378 linksets between them.

Table 1
Generated fingerprint for the rkb-explorer-newcastle dataset.

Category	**value**	Category	**value**
Agriculture	0	Humanities	0.20
Applied Science	0.02	Language	0.08
Arts	0.06	Law	0.01
Belief	0.23	life	0.10
Business	0.22	Mathematics	0.14
Chronology	0.20	Nature	0.21
Culture	0.21	People	0
Education	0.22	Politics	0.06
Environment	0.03	Science	0.39
Geography	0.04	Society	0.51
Health	0	Technology	0.47
History	0	-	-

As in [1,2,4,11,12], linksets were used to define the gold standard for the dataset interlinking recommendation approach implemented in the DRX tool. That is, the evaluation consisted in removing the existing linksets between datasets and verifying to what extent DRX was able to include known interlinked datasets in the recommendation lists it produces. The performance of DRX was measured using the overall *Mean Average Precision* (MAP), as explained in what follows..

Note that the gold standard comprises only the datasets listed in the Mannheim catalog for which the fingerprints can be computed. We deemed as unsuitable the datasets with no associated data or with inaccessible endpoints, even if their metadata would indicate the existence of linksets. Clearly, there is no reason to recommend a dataset that is not accessible to participate in an interlinking process.

More precisely, let d_t be a *target dataset* for which one wants to recommend datasets to be interlinked with and L_t be a ranked list of datasets recommended for d_t. Let G_{d_t} be the gold standard for d_t, i.e., the set of datasets that have linksets with d_t in the gold standard. A dataset d_j is *relevant* for d_t, in the context of G_{d_t}, iff there are linksets connecting d_j and d_t in G_{d_t}. We then define:

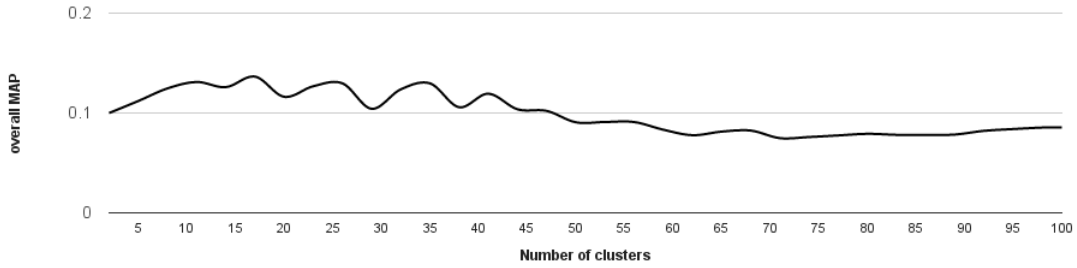

Fig. 3. Strategy 1: Overall Mean Average Precision *vs.* number of clusters.

- $Prec@k\,(L_t)$, the *precision at position* k of L_t, is the number of relevant datasets in L_t until position k.
- $AveP\,(L_t)$ is the average precision at position k of L_t, defined as: $AveP(L_t) = \sum_k Prec@k\,(L_t)\,/\,|G_{d_t}|$.

Recall from Section 2, that the ranked list L_t of datasets recommended for d_t can be generated using two strategies: (i) *cluster-based*, that is, based on the datasets available within a cluster; and (ii) *profiling-based*, that is, based on all datasets available. The *overall MAP* for these strategies is then defined in slightly different ways.

For the profiling-based strategy, we define:

- *The overall MAP* is the average of $AveP\,(L_t)$

and, for the cluster-based strategy, we define:

- MAP (C_i), the *Mean Average Precision* for C_i, is the mean of the average precision at position k of the ranked lists of datasets recommended for the dataset in C_i.
- The *overall MAP* is the average of the MAPs of the clusters.

4.2. Results

We ran experiments considering the two recommendation strategies implemented.

For the cluster-based strategy, Figure 3 shows the the overall MAP as a function of the number of clusters (in increments of 2). It indicates that the maximum value of overall MAP is 13.64%, when the number of clusters was equal to 18.

This result deserves a few comments. First, we note that the overall MAP is obviously influenced by the quality of the gold standard, which in this case may miss possible linksets between datasets. Hence, some recommended datasets may be considered as false positives, when they should have been considered true positives [4].

Second, the maximum MAP is obtained with 18 clusters, whereas the number of categories used to classify datasets in the LOD diagram is only 8. But if we construct just 8 clusters, our recommendation approach reaches an overall MAP of 11.0%, which is sub-optimal. That is, the LOD diagram is not a good starting point for our recommendation strategy. With only 8 clusters, many more non-relevant datasets end up being recommended, which decreases the overall MAP, as compared with the scenario that considers 18 clusters.

For the profiling-based strategy, Figure 4 presents the percentage of the total number of datasets as a function of overall MAP intervals. It shows that this strategy reached an overall MAP of 11-20% for 42% of the datasets. Furthermore, this strategy achieved an overall MAP higher than 20% for more than 30% of the datasets, reaching, in some cases, MAP values higher than 80%.

5. Conclusions and Future Work

In this paper, we proposed a tool, called DRX, to assist data publishers in the process of dataset interlinking and browsing the LOD cloud. A data publisher may use DRX to identify datasets that potentially have resources to be interlinked with a given dataset. DRX takes advantage of various methods including crawling, profiling, clustering and ranking modules to create ranked lists of datasets to be interlinked with a given dataset.

The results obtained indicate that the proposed approach can indeed be used to facilitate the task of dataset interlinking in the LOD. They show that the

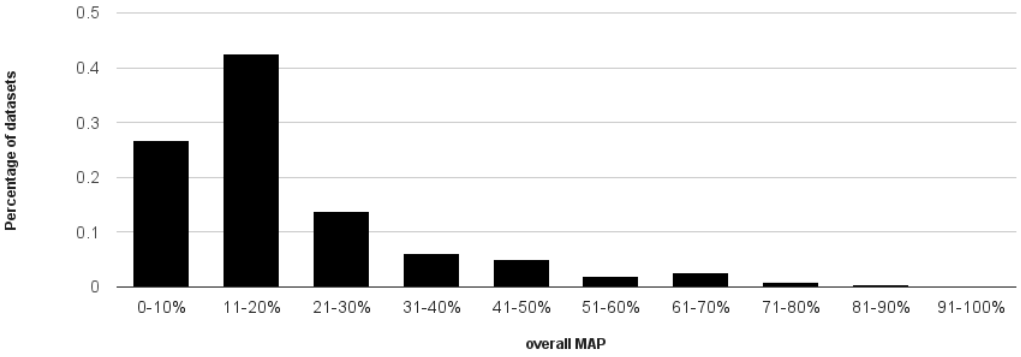

Fig. 4. Strategy 2: Percentage of datasets *vs.* Overall Mean Average Precision.

profiling-based strategy achieves a better performance than the cluster-based strategy.

References

[1] A. A. M. Caraballo, B. P. Nunes, G. R. Lopes, L. A. P. Paes Leme, M. A. Casanova, and S. Dietze, *TRT-A Tripleset Recommendation Tool*, in International Semantic Web Conference (Posters & Demos), pp. 105-108. 2013.

[2] A. A. M. Caraballo, N. M. Arruda Jr, B. P. Nunes, G. R. Lopes, and M. A. Casanova, *TRTML-A Tripleset Recommendation Tool Based on Supervised Learning Algorithms*, in The Semantic Web: ESWC 2014 Satellite Events, pp. 413-417, Springer International Publishing, 2014.

[3] 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.

[4] M. Emaldi, O. Corcho, and D. López-de-Ipiña, *Detection of Related Semantic Datasets Based on Frequent Subgraph Mining Mikel*, in Proceedings of the 4th International Workshop on Intelligent Exploration of Semantic Data (IESD) 2015, 14th International Semantic Web Conference ISWC, 2015.

[5] B. Fetahu, S. Dietze, B. P. Nunes, M. A. Casanova, D. Taibi and W. Nejdl, *A scalable approach for efficiently generating structured dataset topic profiles*, in The Semantic Web: Trends and Challenges, pp. 519-534. Springer International Publishing, 2014.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *WEKA data mining software: an update*, ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18, 2009.

[7] T. Heath, and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space* (1st Edition), volume 1 of Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011. Available from `http://linkeddatabook.com/editions/1.0/`.

[8] A. Jentzsch, R. Cyganiak, C. Bizer, *State of the lod cloud*, 2011.

[9] R. Kawase, P. Siehndel, B. P. Nunes, E. Herder and W. Nejdl, *Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources*, in Proceedings of the 25th ACM conference on Hypertext and social media, pp. 56-65. ACM, 2014.

[10] S. Lalithsena, P. Hitzler, A. Sheth and Paril Jain, *Automatic domain identification for linked open data*, in Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, vol. 1, pp. 205-212. IEEE, 2013.

[11] L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze, *Identifying candidate datasets for data interlinking*, in Web Engineering, pp. 354-366. Springer Berlin Heidelberg, 2013.

[12] G. R. Lopes, L. A. P. Paes Leme, B. P. Nunes, M. A. Casanova, and S. Dietze,*Recommending tripleset interlinking through a social network approach*, in Web Information Systems Engineering, WISE 2013, pp. 149-161. Springer Berlin Heidelberg, 2013.

[13] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. *DBpedia spotlight: shedding light on the web of documents.* In Proceedings of the 7th International Conference on Semantic Systems, pp. 1-8. ACM, 2011.

[14] A. Nikolov, and M. d'Aquin, *Identifying relevant sources for data linking using a semantic web index*, in: Proceedings of the Linked Data on the Web workshop in conjunction with the 20th international World Wide Web conference, vol. 813 (2011).

[15] A. Ngomo, and S. Auer, *LIMES: a time-efficient approach for large-scale link discovery on the web of data.* In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. pp. 2312-2317 (2011).

[16] D. Pelleg and A. W. Moore, *X-means: Extending K-means with Efficient Estimation of the Number of Clusters*, in ICML, pp. 727-734, 2000.

[17] M. Schmachtenberg, C. Bizer, and H. Paulheim, *Adoption of the linked data best practices in different topical domains.* In The Semantic Web-ISWC 2014, pp. 245-260. Springer International Publishing, 2014.

[18] B., Sergey, and L. Page. *Reprint of: The anatomy of a large-scale hypertextual web search engine.* Computer networks 56, no. 18 (2012): 3825-3833.