Dataset Recommendation for Data Linking: An Intensional Approach

Mohamed Ben Ellefi, Zohra Bellahsene, Konstantin Todorov, Stefan Dietze

Dataset Recommendation for Data Linking: an Intensional Approach

Mohamed Ben Ellefi[1], Zohra Bellahsene[1], Stefan Dietze[2], Konstantin Todorov[1]

[1]LIRMM / University of Montpellier, France
{benellefi, bella, todorov}@lirmm.fr
[2]L3S Research Center / Leibniz University Hannover, Germany
dietze@l3s.de

Abstract. With the growing quantity and diversity of publicly available web datasets, most notably Linked Open Data, recommending datasets, which meet specific criteria, has become an increasingly important, yet challenging problem. This task is of particular interest when addressing issues such as entity retrieval, semantic search and data linking. Here, we focus on that last issue. We introduce a dataset recommendation approach to identify linking candidates based on the presence of schema overlap between datasets. While an understanding of the nature of the content of specific datasets is a crucial prerequisite, we adopt the notion of dataset profiles, where a dataset is characterized through a set of schema concept labels that best describe it and can be potentially enriched by retrieving their textual descriptions. We identify schema overlap by the help of a semantico-frequential concept similarity measure and a ranking criterium based on the *tf*idf* cosine similarity. The experiments, conducted over all available linked datasets on the Linked Open Data cloud, show that our method achieves an average precision of up to 53% for a recall of 100%. As an additional contribution, our method returns the mappings between the schema concepts across datasets – a particularly useful input for the data linking step.

1 Introduction

With the emergence of the Web of Data, in particular Linked Open Data (LOD) [1], an abundance of data has become available on the web. Dataset recommendation is becoming an increasingly important task to support challenges such as entity interlinking [2], entity retrieval or semantic search [3]. Particularly with respect to interlinking, the current topology of the LOD cloud underlines the need for practical and efficient means to recommend suitable datasets: currently, only very few, well established knowledge graphs show a high amount of inlinks, with DBpedia being the most obvious target [4], while a large amount of datasets is largely ignored.

This is due in part to the challenge to identify suitable linking candidates without prior knowledge of the available datasets and their characteristis. Linked datasets vary significantly with respect to represented resource types, currentness, coverage of topics and domains, size, used languages, coherence, accessibility [5] or general quality aspects [6]. This heterogeneity poses significant challenges for data consumers when attempting to find useful datasets. Hence, a long tail of datasets from the LOD cloud[1] has hardly been reused and adopted, while the majority of data consumption, linking and reuse focuses on established knowledge graphs such as DBpedia [7] or YAGO [8].

In line with [9], we define dataset recommendation as the problem of computing a rank score for each of a set of datasets D_T (for Target Dataset) so that the rank score indicates the relatedness of D_T to a given dataset, D_S (for Source Dataset). The rank scores provide information of the likelihood of a D_T dataset to contain linking candidates for D_S.

We adopt the notion of a dataset profile, defined as a set of concept labels that describe the dataset. By retrieving the textual descriptions of each of these labels, we can map the label profiles to larger text documents. This representation provides richer contextual and semantic information and allows to compute efficiently and inexpensively similarities between profiles.

Although different types of links can be defined across datasets, here we focus on the identity relation given by the statement “owl:sameAs”. Our working hypothesis is simple: datasets that share at least one concept, i.e., at least one pair of semantically similar concept labels, are likely to contain at least one potential pair of instances to be linked by a “owl:sameAs” statement. We base our recommendation procedure on this hypothesis and propose an approach in two steps: (1) for every D_S, we identify a cluster[2] of datasets that share schema concepts with D_S and (2) we rank the datasets in each cluster with respect to their relevance to D_S.

In step (1), we identify concept labels that are semantically similar by using a similarity measure based on the frequency of term co-occurence in a large corpus (the web) combined with a semantic distance based on WordNet without relying on string matching techniques [10]. For example, this allows to recommend to a dataset annotated by “school” one annotated by “college”. In this way, we form clusters of “comparable datasets” for each source dataset. The intuition is that for a given source dataset, any of the datasets in its cluster is a potential target dataset for interlinking.

Step (2) focuses on ranking the datasets in a D_S-cluster with respect to their importance to D_S. This allows to evaluate the results in a more meaningful way and of course to provide quality results to the user. The ranking criterium should not be based on the amount of schema overlap, because potential to-link instances can be found in datasets sharing 1 class or sharing 100. Therefore, we need a similarity measure on the *profiles* of the comparable datasets. We have proceeded by building a vector model for the document representations of the profiles and computing cosine similarities.

[1] `http://datahub.io/group/lodcloud`

[2] We note that we use the term ”cluster” in its general meaning, referring to a set of datasets grouped together by their similarity and not in a machine learning sense.

To evaluate the approach, we have used the current topology of the LOD as evaluation data (ED). As mentioned in the beginning, the LOD link graph is far from being complete, which complicates the interpretation of the obtained results—many false positives are in fact missing positives (missing links) from the evaluation data—a problem that we discuss in detail in the sequel. Note that as a result of the recommendation process, the user is not only given candidate datasets for linking, but also pairs of classes where to look for identical instances. This is an important advantage allowing to run more easily linking systems like SILK [11] in order to verify the quality of the recommendation and perform the acutal linking. Our experimental tests with SILK confirm the hypothesis on the incompleteness of the ED.

To sum up, the paper contains the following contributions: (1) new definitions of dataset profiles based on schema concepts, (2) a recommendation framework allowing to identify the datasets sharing schema with a given source dataset, (3) an efficient ranking criterium for these datasets, (4) an output of additional metadata such as pairs of similar concepts across source and target datasets, (5) a large range of reproducible experiments and in depth analysis with all of our results made available.

We proceed to present the theoretical grounds of our technique in Section 2. Section 3 defines the evaluation framework that has been established and reports on our experimental results. Related approaches are presented and discussed in Section 4 before we conclude in Section 5.

2 A Dataset Interlinking Recommendation Framework

Our recommendation approach relies on the notion of a dataset profile, providing comparable representations of the datasets by the help of characteristic features. In this section, we first introduce the definitions of a dataset profile that we are using in this study. Afterwards, we describe the profile-based recommendation technique that we apply.

2.1 Intensional Dataset Profiles

A dataset profile is seen as a set of dataset characteristics that allow to describe in the best possible way a dataset and that separate it maximally from other datasets. A feature-based representation of this kind allows to compute distances or measure similarities between datasets (or for that matter profiles), which unlocks the dataset recommendation procedure. These descriptive characteristics, or features, can be of various kinds (statistical, semantic, extensional, etc.). As we observe in [12], a dataset profile can be defined based on a set of types (schema concepts) names that represent the topic of the data and the covered domain. In line with that definition, we are interested here in intensional dataset characteristics in the form of a set of keywords together with their definitions that best describe a dataset.

Definition 1 (Dataset Label Profile). *The* **label profile** *of a dataset D, denoted by $\mathcal{P}_l(D)$, is defined as the set of n schema concept labels corresponding to D: $\mathcal{P}_l(D) = \{L_i\}_{i=1}^n$.*

Note that the representativity of the labels in $\mathcal{P}_l(D)$ with respect to D can be improved by filtering out certain types. We rely on two main heuristics: (1) remove too popular types (such as foaf:Person), (2) remove types with too few instances in a dataset. These two heuristics are based on the intuition that the probability of finding identical instances of very popular or underpopulated classes is low. We support (1) experimentally in Section 3 while we leave (2) for future work.

Each of the concept labels in $\mathcal{P}_l(D)$ can be mapped to a text document consisting of the label itself and a textual description of this label. This textual description can be the definition of the concept in its ontology, or any other external textual description of the terms composing the concept label. We define a document profile of a dataset in the following way.

Definition 2 (Dataset Document Profile). *The* **document profile** *of a dataset D, $\mathcal{P}_d(D)$, is defined as a text document constructed by the concatenation of the labels in $\mathcal{P}_l(D)$ and the textual descriptions of the labels in $\mathcal{P}_l(D)$.*

Note that there is no substantial difference between the two definitions given above. The document profile is an extended label profile, where more terms, coming from the label descriptions, are included. This allows to project the profile similarity problem onto a vector space by indexing the documents and using a term weighting scheme of some kind (e.g., *tf*idf*).

By the help of these two definitions, a profile can be constructed for any given dataset in a simple and inexpensive way, independent on its connectivity properties on the LOD. In other words, a profile can be easily computed for datasets that are already published and linked, just as for datasets that are to be published and linked, allowing to use the same representation for both kinds of datasets and thus allowing for their comparison by the help of feature-based similarity measures.

As stated in the introduction, we rely on the simple intuition that datasets with similar intension have extensional overlap. Therefore, it suffices to identify at least one pair of semantically similar types in the schema of two datasets in order to select these datasets as potential linking candidates. We are interested in the semantic similarity of concept labels in the dataset label profiles. There are many off-the-shelf similarity measures that can be applied, known from the ontology matching literature. We have focused on the well known semantic measures Wu Palmer [13] and Lin's [14], as well as the UMBC [10] measure that combines semantic distance in WordNet with frequency of occurrence and co-occurrence of terms in a large external corpus (the web). We provide the definition of that measure, since it is less well-known and it showed to perform best in our experiments. For two labels, x and y, we have

$$sim_{\mathrm{UMBC}}(x,y) = sim_{\mathrm{LSA}}(x,y) + 0.5e^{-\alpha D(x,y)}, \tag{1}$$

where $sim_{\mathrm{LSA}}(x, y)$ is the Latent Semantic Analysis (LSA) [15] word similarity, which relies on the words co-occurrence in the same contexts computed in a three billion words corpus[3] of good quality English. $D(x, y)$ is the minimal WordNet [16] path length between x and y. According to [10], using $e^{-\alpha D(x,y)}$ to transform simple shortest path length has shown to be very efficient when the parameter α is set to 0.25.

With a concept label similarity measure at hand, we introduce the notion of dataset comparability, based on the existence of shared intension.

Definition 3 (Comparable Datasets). *Two datasets D' and D'' are comparable if there exists L_i and L_j such that $L_i \in \mathcal{P}_l(D')$, $L_j \in \mathcal{P}_l(D'')$ and $sim_{UMBC}(L_i, L_j) \geq \theta$, where $\theta \in [0, 1]$.*

2.2 Recommendation Process: the CCD-CosineRank approach

A dataset recommendation procedure for the linking task returns, for a given source dataset, a set of target datasets ordered by their likelihood to contain instances identical to those in the source dataset.

Let D_S be a source dataset. We introduce the notion of a *cluster of comparable datasets* related to D_S, or $CCD(D_S)$ for short, defined as the set of target datasets, denoted by D_T, that are comparable to D_S according to Def. 3. Thus, D_S is identified by its CCD and all the linking candidates D_T for this dataset are found in its cluster, following our working hypothesis.

Finally, we need a ranking function that assigns scores to the datasets in $CCD(D_S)$ with respect to D_S expressing the likelihood of a dataset in $CCD(D_S)$ to contain identical instances with those of D_S. To this end, we need a similarity measure on the dataset profiles.

We have worked with the document profiles of the datasets (Def. 2). Since datasets are represented as text documents, we can easily build a vector model by indexing the documents in the corpus formed by all datasets of interest – the ones contained in one single CCD. We use a *tf*idf* weighting scheme, which allows to compute the cosine similarity between the document vectors and thus assign a ranking score to the datasets in a CCD with respect to a given dataset from the same CCD. Note that this approach allows to consider the information of the intensional overlap between datasets prior to ranking and indexing – we are certain to work only with potential linking candidates when we rank, which improves the quality of the ranks. For a given dataset D_S, the procedure returns datasets from $CCD(D_S)$, ordered by their cosine similarity to D_S.

Finally, an important outcome of the recommendation procedure is the fact that, along with an ordered list of linking candidates, the user is provided the pairs of types of two datasets—a source and a target—where to look for identical instances. This information facilitates considerably the linking process, to be performed by an instance matching tool, such as SILK.

[3] http://ebiquity.umbc.edu/resource/html/id/351

2.3 Application of the Approach: an Example

We illustrate our approach by an example. We consider *education-data-gov-uk*[4] as a source dataset (D_S). The first step consists in retrieving the schema concepts from this dataset and constructing a clean label profile (we filter out noisy labels, as discussed above), as well as its corresponding document profile (Def. 1 and Def. 2, respectively). We have $\mathcal{P}_l(\textit{education-data-gov-uk}) = \{$London Borough Ward, School, Local Learning Skills Council, Adress$\}$. We perform a semantic comparison between the labels in $\mathcal{P}_l(\textit{education-data-gov-uk})$ and all labels in the profiles of the accessible LOD datasets. By fixing $\theta = 0.7$, we generate $CCD(\textit{education-data-gov-uk})$ containing the set of comparable datasets D_T, as described in Def. 3. The second step consists of ranking the D_T datasets in $CCD(\textit{education-data-gov-uk})$ by computing the cosine similarity between their document profiles and $\mathcal{P}_d(\textit{education-data-gov-uk})$. The top 5 ranked candidate datasets to be linked with *education-data-gov-uk* are (1) *rkb-explorer-courseware*[5], (2) *rkb-explorer- courseware*[6], (3) *rkb-explorer-southampton*[7], (4) *rkb-explorer-darmstadt*[8], and (5) *oxpoints*[9].

Finally, for each of these datasets, we retrieve the pairs of shared (similar) schema concepts extracted in the comparison part:

– *education-data-gov-uk* and *statistics-data-gov-uk* share two labels "London Borough Ward" and "LocalLearningSkillsCouncil".

– *education-data-gov-uk* and *oxpoints* contain similar labels which are, respectively, "School" and "College", for the SILK results see Section 3.5.

3 Experiments and Results

We proceed to report on the experiments conducted in support of the proposed recommendation method.

3.1 Evaluation Framework

The quality of the outcome of a recommendation process can be evaluated along a number of dimensions. Ricci *et al.* [17] provide a large review of recommender systems evaluation techniques and cite three common types of experiments: (i) offline experiments, where recommendation approaches are compared without user interaction, (ii) user studies, where a small group of subjects experiments with the system and reports on the experience, and (iii) online experiments, where real user populations interact with the system.

For the task of dataset recommendation, the system suggests to the user a list of n target datasets candidates to be linked to a given source dataset. There

[4] `http://education.data.gov.uk/`

[5] `http://courseware.rkbexplorer.com/`

[6] `http://courseware.rkbexplorer.com/`

[7] `http://southampton.rkbexplorer.com/`

[8] `http://darmstadt.rkbexplorer.com/`

[9] `https://data.ox.ac.uk/sparql/`

does not exist a common evaluation framework for the datasets recommendation, thus, we evaluate our method with an offline experiment by using a pre-collected set of linked data considered as evaluation data (ED). The most straightforward, although not unproblematic (see below) choice of evaluation data for the data linking recommendation task is the existing link topology of the current version of the LOD cloud.

In our recommendation process, for a given source dataset D_S, we identify a cluster of target datasets, D_T, that we rank with respect to D_S (*cf.* Section 2.2). To evaluate the quality of the recommendation results given the ED of our choice, we compute the common evaluation measures for recommender systems, precision and recall, defined as functions of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) as follows:

$$Pr = \frac{TP}{TP + FP}; \quad Re = \frac{TP}{TP + FN}. \tag{2}$$

The number of potentially useful results that can be presented to the user has to be limited. Therefore, to assess the effectiveness of our approach, we rely on the measure of precision at rank k denoted by $P@k$. Complementarily, we evaluate the precision of our recommendation when the level of recall is 100% by using the mean average precision at $Recall = 1$, MAP@R, given as:

$$MAP@R = \frac{\sum_{q=1}^{\text{Total}_{D_S}} Pr@R(q)}{\text{Total}_{D_S}}, \tag{3}$$

where $R(q)$ corresponds to the rank, at which recall reaches 1 for the qth dataset and Total_{D_S} is the entire number of source datasets in the evaluation.

3.2 Experimental Setup

We started by crawling all available datasets in the LOD cloud group on the Data Hub[10] in order to extract their profiles. In this crawl, only 90 datasets were accessible via endpoints or via dump files. In the first place, for each accessible dataset, we extracted its implicit and explicit schema concepts and their labels, as described in Def. 1. The explicit schema concepts are provided by resource types, while the implicit schema concepts are provided by the definitions of a resource properties [18]. As noted in Section 2, some labels such as “Group”, “Thing”, “Agent”, “Person” are very generic, so they are considered as noisy labels. To address this problem, we filter out schema concepts described by generic vocabularies such as VoID[11], FOAF[12] and SKOS[13]. The dataset document profiles, as defined in Def. 2, are constructed by extracting the textual descriptions

[10] `http://datahub.io/group/lodcloud`

[11] `http://rdfs.org/ns/void`

[12] `http://xmlns.com/foaf/0.1/`

[13] `http://www.w3.org/2004/02/skos/core`

of labels by querying the Linked Open Vocabularies[14] (LOV) with each of the concept labels per dataset.

To form the clusters of comparable datasets from Def. 3, we compute the semantico-frequential similarity between labels (given in eq. (1)). We apply this measure via its available web API service[15]. In addition, we tested our system with two more semantic similarity measures based on WordNet: Wu Palmer and Lin's. For this purpose, we used the 2013 version of the $WS4J$[16] java API.

The evaluation data (ED) corresponds to the outgoing and incoming links extracted from the generated VoID file using the *datahub2void* tool[17]. It is made available on `http://www.lirmm.fr/benellefi/void.ttl`. We note that out of 90 accessible datasets, only those that are linked to at least one accessible dataset in the ED are evaluated in the experiments.

3.3 Evaluation Results

We started by considering each dataset in the ED as an unlinked source (newly published) dataset D_S. Then, we ran the *CCD-CosineRank* workflow, as described in Section 2.2. The first step is to form a $CCD(D_S)$ for each D_S. The CCD construction process depends on the similarity measure on dataset profiles. Thus, we evaluated the CCD clusters in terms of recall for different levels of the threshold θ (*cf.* Def. 3) for the three similarity measures that we apply. We observed that the recall value remains 100% in the following threshold intervals per similarity measure: **Wu Palmer**: $\theta \in [0, 0.9]$; **Lin**: $\theta \in [0, 0.8]$; **UMBC**: $\theta \in [0, 0.7]$.

The CCD construction step ensures a recall of 100% for various threshold values, which will be used to evaluate the ranking step of our recommendation process by the Mean Average Precision (MAP@R) at the maximal recall level, as defined in Def. 3. The results in Fig. 1 show highest performance of the UMBC's measure with a $MAP@R \cong 53\%$ for $\theta = 0.7$, while the best MAP@R values for Wu Palmer and Lin's measures are, respectively, 50% for $\theta = 0.9$ and 51% for $\theta = 0.8$. Guided by these observations, we evaluated our ranking in terms of precision at ranks $k = \{5, 10, 15, 20\}$, as shown in Table 1. Based on these results, we choose UMBC at a threshold $\theta = 0.7$ as a default setting for *CCD-CosineRank*, since it performs best for three out of four k-values and it is more stable than the two others especially with MAP@R.

3.4 Baselines and Comparison

To the best of our knowledge, there does not exist a common benchmark for dataset interlinking recommendation. Since our method uses both label profiles and document profiles, we implemented two recommendation approaches to be

[14] `http://lov.okfn.org/dataset/lov/`

[15] `http://swoogle.umbc.edu/SimService/`

[16] `https://code.google.com/p/ws4j/`

[17] `https://github.com/lod-cloud/datahub2void`

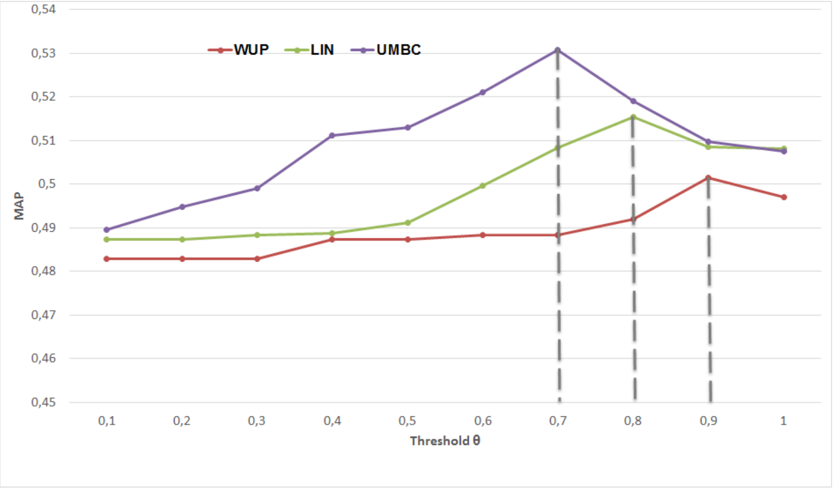

Fig. 1. The MAP@R of our recommender system by using three different similarity measures for different similarity threshold values

considered as baselines – one using document profiles only, and another one using label profiles:

***Doc-CosineRank*:** All datasets are represented by their *document profiles*, as given in Def. 2. We build a vector model by indexing the documents in the corpus formed by all available LOD datasets (no CCD clusters). We use a *tf*idf* weighting scheme, which allows us to compute the cosine similarity between the document vectors and thus assign a ranking score to each dataset in the entire corpus with respect to a given dataset D_S.

UMBCLabelRank: All datasets are represented by their *label profiles*, as given in Def. 1. For a source dataset D_S, we construct its $CCD(D_S)$ according to Def. 3 using UMBC with $\theta = 0.7$. Thus, D_S is identified by its CCD and all target datasets D_T are found in its cluster. Let AvgUMBC be a ranking function that assigns scores to each D_T in $CCD(DS)$, defined by:

$$Avg\text{UMBC}(D', D'') = \frac{\sum_{i=1}^{|\mathcal{P}_l(D')|} \sum_{j=1}^{|\mathcal{P}_l(D'')|} \max sim_{UMBC}(L_i, L_j)}{\max(|\mathcal{P}_l(D')|, |\mathcal{P}_l(D'')|)}, \qquad (4)$$

where L_i in $\mathcal{P}_l(D')$ and L_j in $\mathcal{P}_l(D'')$.

Fig. 2 depicts a detailed comparison of the precisions at recall 1 obtained by the three approaches for each D_S taken as source dataset. It can be seen that the *CCD-CosineRank* approach is more stable and largely outperforms the two other approaches by an MAP@R of up to 53% as compared to 39% for *UMBCLabelRank* and 49% for *CCD-CosineRank*. However, the *UMBCLabelRank* approach produces better results than the other ones for a limited number of

Measure \ P@k	P@5	P@10	P@15	P@20
WU Palmer ($\theta = 0.9$)	$0,56$	$0,52$	0.53	0.51
Lin ($\theta = 0.8$)	0.57	**0.54**	**0.55**	0.51
UMBC ($\theta = 0.7$)	**0.58**	**0.54**	0.53	**0.53**

Table 1. Precision at 5, 10, 15 and 20 of the *CCD-CosineRank* approach using three different similarity measures over their best threshold values based on Fig.1

source datasets, especially in the case when D_S and D_T share a high number of identic labels in their profiles.

The performance of the *CCD-CosineRank* approach demonstrates the efficiency and the complementarity of combining in the same pipeline (i) the *semantic* similarity on labels for identifying recommendation candidates (*CCD* construction process) and (ii) the *frequential* document cosine similarity to rank the candidate datasets. We make all of the ranking results of the *CCD-CosineRank* approach available to the community on `http://www.lirmm.fr/benellefi/CCD-CosineRank_Result.csv`.

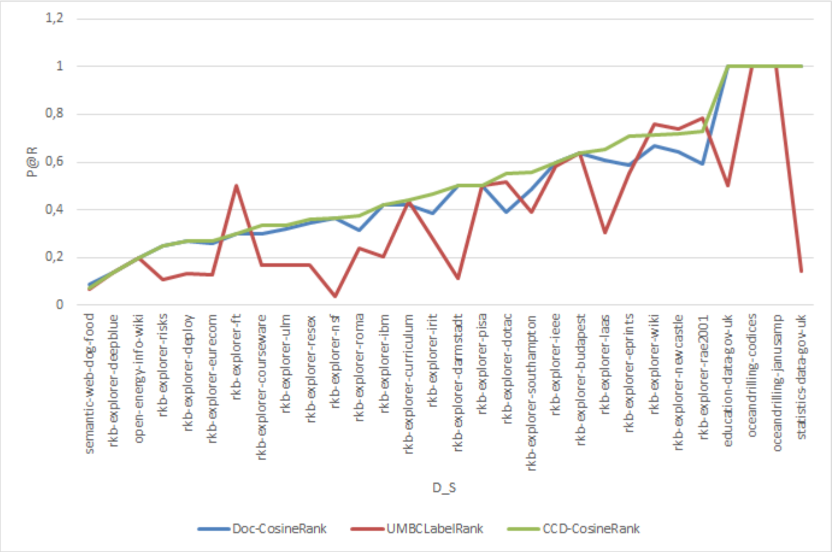

Fig. 2. Precisions at recall=1 of the *CCD-CosineRank* approach as compared to *Doc-CosineRank* and *UMBCLabelRank*

3.5 Discussion

We begin by a note on the vocabulary filtering that we perform (Section. 3.2). We underline that we have identified the types which improve/decrease the performance empirically. As expected, vocabularies, which are very generic and wide-spread have a negative impact, acting like hub nodes, which dilute the results. For comparison, the results of the recommendation before removal are made available on `http://www.lirmm.fr/benellefi/RankNoFilter.csv`.

The different experiments described above show a high performance of the introduced recommendation approach with an average precision of 53% for a recall of 100%. Likewise, it may be observed that this performance is completely independent of the dataset size (number of triples) or the schema cardinality (number of schema concepts by datasets). However, we note that better performance was obtained for datasets from the *geographic* and *governmental* domains with precision and recall of 100%. Naturally, this is due to the fact that a recommender system in general and particularly our system performs better with datasets having high quality schema description and datasets reusing existing vocabularies (the case for the two domains cited above), which is considered as linked data modeling best practice. An effort has to be made for improving the quality of the published dataset [19].

We believe that our method can be given a more fair evaluation if better evaluation data in the form of ground truth are used. Indeed, our results are impacted by the problem of false positives overestimation. Since data are not collected using the recommender system under evaluation, we are forced to assume that the false positive items would have not been used even if they had been recommended, i.e., that they are uninteresting or useless to the user. This assumption is, however, generally false, for example when the set of unused items contains some interesting items that the user did not select. In our case, we are using declared links in the LOD cloud as ED, which is certain but far from being complete for it to be considered as ground truth. Thus, in the recommendation process the number of false positives tends to be overestimated, or in other words an important number of missing positives in the ED translates into false positives in the recommendation process.

To further illustrate the effect of false positives overestimation, we ran SILK as an instance matching tool to discover links between D_S and their corresponding D_Ts that have been considered as false positives in our ED. SILK takes as an input a *Link Specification Language* file, which contains the instance matching configuration. We recall that our recommendation procedure provides pairs of shared or similar types between D_S and every D_T in its corresponding CCD, which are particularly useful to configure SILK. However, all additional information, such as the datatype properties of interest, has to be given manually. This makes the process very time consuming and tedious to perform over the entire LOD. Therefore, as an illustration, we ran the instance matching tool on two flagship examples of false positive D_Ts:

Semantically Similar Labels: We choose *education-data-gov-uk*[18] as a D_S and its corresponding false positive D_T *oxpoints*[19]. The two datasets contain

[18] `http://education.data.gov.uk/`

[19] `https://data.ox.ac.uk/sparql`

in their profiles, respectively, the labels "School" and "College", detected as highly similar labels by the UMBC measure, with a score of 0.91. The instance matching gave as a result 10 accepted "owl:sameAs" links between the two datasets.

Identical Labels: We choose *rkb-explorer-unlocode*[20] as a D_S and its corresponding D_Ts, which are considered as FP: *yovisto*[21] *datos-bcn-uk*[22] *datos-bcn-cl*[23]. All 4 datasets share the label "Country" in their corresponding profiles. The instance matching process gave as a result a set of accepted "owl:sameAs" links between *rkb-explorer-unlocode* and each of the three D_T.

We provide the set of newly discovered linksets to be added to the LOD topology and we made the generated linksets and the corresponding SILK configurations available on `http://www.lirmm.fr/benellefi/Silk_Matching`.

It should be noted that the recommendation results provided by our approach may contain some broader candidate datasets with respect to the source dataset. For example, two datasets that share schema labels such as books and authors are considered as candidates even when they are from different domains like science vs. literature. This outcome can be useful for predicting links such as "rdfs:seeAlso" (rather than "owl:sameAs"). We have chosen to avoid the inclusion of instance-related information in order to keep the complexity of the system as low as possible and still provide reasonable precision by guaranteeing a 100% recall.

As a conclusion, we outline three directions of work in terms of dataset quality that can considerably facilitate the evaluation of any recommender system in that field: (1) improving descriptions and metadata; (2) improving accessibility; (3) providing a reliable ground truth and benchmark data for evaluation.

4 Related Work

With respect to finding relevant datasets on the Web, we cite briefly several studies on discovering relevant datasets for query answering. Based on well-known data mining strategies, [20] and [21] present techniques to find relevant datasets, which offer contextual information corresponding to the user queries. A feedback-based approach to incrementally identify new datasets for domain-specific linked data applications is proposed in [22]. User feedback is used as a way to assess the relevance of the candidate datasets.

In the following, we cite approaches that have been devised for the datasets interlinking candidates recommendation task and which are, therefore, directly relevant to our work. Nikolov *et al.* [23] propose a keyword-based search approach to identify candidate sources for data linking consisting of two steps: (i) searching for potentially relevant entities in other datasets using as keywords

[20] `http://unlocode.rkbexplorer.com/sparql/`

[21] `http://sparql.yovisto.com/`

[22] `http://data.open.ac.uk/query`

[23] `http://data.open.ac.cl/query`

randomly selected instances over the literals in the source dataset, and (ii) filtering out irrelevant datasets by measuring semantic concept similarities obtained by applying ontology matching techniques.

Mehdi *et al.* [24] propose a method to automatically identify relevant public SPARQL endpoints from a list of candidates. First, the process needs as input a set of domain-specific keywords, which are extracted from a local source or can be provided manually by an expert. Then, using natural languages processing techniques and queries expansion techniques, the system generates a set of queries that seek to exact literal matches between the introduced keywords and the target datasets, i.e., for each term supplied to the algorithm, the system runs a comparison to a set of eight queries: {original-case, proper-case, lower-case, upper-case} × {no-lang-tag, @en-tag}. Finally, the produced output consists of a list of potentially relevant SPARQL endpoints of datasets for linking. In addition, an interesting contribution of this technique is the bindings returned for the subject and predicate query variables, which are recorded and logged when a term match is found on some particular SPARQL endpoint. The records are useful in the linking step.

Leme *et al.* [25] present a ranking method for datasets with respect to their relevance for the interlinking task. The ranking is based on Bayesian criteria and on the popularity of the datasets, which affects the generality of the approach. The authors extend this work and overcome this drawback in [9] by exploring the correlation between different sets of features—properties, classes and vocabularies—and the links to compute new rank score functions for all the available linked datasets.

None of the studies outlined above have evaluated the ranking measure in terms of Precision/Recall, except for [9] which, according to the authors, achieves a mean average precision of around 60% and an excepted recall of 100% with rankings over all LOD datasets. However, a direct comparison to our approach seems unfair since the authors did not provide the list of the datasets and their rank performance by datasets considered as source.

In comparison to the work discussed above, our approach has the potential of overcoming a series of complexity related problems, precisely, considering the complexity to generate the matching in [23], to produce the set of domain-specific keywords as input in [24] and to explore the set of features of all the network datasets in [9]. Our recommendation results are much easier to obtain since we only manipulate the schema part of the dataset. They are also easier to interpret and apply since we automatically recommend the corresponding schema concept mappings together with the candidate datasets.

5 Conclusion and Future Work

Following the linked data best practices, metadata designers reuse and build on, instead of replicating, existing RDF schema and vocabularies. Motivated by this observation, we propose the *CCD-CosineRank* interlinking candidate dataset recommendation approach, based on concept label profiles and schema overlap across datasets. Our approach consists of identifying clusters of comparable datasets, then, ranking the datasets in each cluster with respect to a given dataset. We discuss three different similarity measures, by which the relevance of our recommendation can be achieved. We evaluate our approach on real data coming from the LOD cloud and compare it two baseline methods. The results show that our method achieves a mean average precision of around 53% for recall of 100%, which reduces considerably the cost of dataset interlinking. In addition, as a post-processing step, our system returns sets of schema concept mappings between source and target datasets, which decreases considerably the interlinking effort and allows to verify explicitly the quality of the recommendation.

In the future, we plan to improve the evaluation framework by developing a more reliable and complete evaluation data for dataset recommendation. We plan to elaborate a ground truth based on certain parts of the LOD, possibly by using crowdsourcing techniques, in order to deal with the false positives overestimation problem. Further work should go into btaining high quality profiles, in particular by considering the population of the schema elements. We also plan to investigate the effectiveness of machine learning techniques, such as classification or clustering, for the recommendaiton task. One of the conclusions of our study shows that the recommendation approach is limited by the lack of accessibility, explicit metadata and quality descriptions of the datasets. As this can be given as an advice to data publishers, in the future, we will work on the development of recommendation methods for datasets with noisy and incomplete descriptions.

Acknowledgements

This research has been partially funded under the Datalyse project[24]-FSN-AAP Big Data n3-, by the European Commission-funded DURAARK project (FP7 Grant Agreement No. 600908) and the COST Action IC1302 (KEYSTONE).

References

1. C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.
2. 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 *Proc. of the 10th ESWC*, pp. 548–562, 2013.
3. R. Blanco, P. Mika, and S. Vigna, “Effective and efficient entity search in rdf data.,” in *Proc. of ISWC*, vol. 7031, pp. 83–97, Springer, 2011.
4. M. Schmachtenberg, C. Bizer, and H. Paulheim, “Adoption of the linked data best practices in different topical domains,” in *Proc. of ISWC*, pp. 245–260, 2014.
5. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbusshe, “Sparql web-querying infrastructure: Ready for action?,” in *Proc. of the 12th ISWC*, 2013.
6. C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann, “Assessing linked data mappings using network measures,” in *Proc. of the 9th ESWC*, pp. 87–102, 2012.

[24] http://www.datalyse.fr/

7. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *proc. of ISWC*, pp. 722–735, 2007.
8. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proc. of WWW*, pp. 697–706, 2007.
9. G. Lopes, L. A. Paes Leme, B. Nunes, M. Casanova, and S. Dietze, "Two approaches to the dataset interlinking recommendation problem," in *Proc. of 15th on WISE 2014*, 2014.
10. L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc_ebiquity-core: Semantic textual similarity systems," in *Proc. of the *SEM*, Association for Computational Linguistics, 2013.
11. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk - A link discovery framework for the web of data," in *Proceedings of the WWW2009, LDOW*, 2009.
12. M. B. Ellefi, Z. Bellahsene, F. Scharffe, and K. Todorov, "Towards semantic dataset profiling," in *Proc. of Dataset PROFIling & fEderated Search for Linked Data Workshop co-located with the 11th ESWC*, 2014.
13. Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. of the 32Nd ACL*, pp. 133–138, 1994.
14. D. Lin, "An information-theoretic definition of similarity," in *Proc. of ICML*, pp. 296–304, 1998.
15. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS1990*, vol. 41, pp. 391–407.
16. G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.
17. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, *Recommender systems handbook*, vol. 1. Springer, 2011.
18. T. Gottron, M. Knauf, S. Scheglmann, and A. Scherp, "A systematic investigation of explicit and implicit schema information on the linked open data cloud," in *Proc. of ESWC*, pp. 228–242, 2013.
19. M. B. Ellefi, Z. Bellahsene, and K. Todorov, "Datavore: A vocabulary recommender tool assisting linked data modeling," in *Proc.of the ISWC Posters & Demonstrations Track a track*, 2015.
20. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Discovering related data sources in data-portals," in *Proc. of the 1st IWSS*, 2013.
21. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Entity-based data source contextualization for searching the web of data," in *Proc. of the Dataset PROFIling & fEderated Search for Linked Data Workshop co-located with the 11th ESWC*, pp. 25–41, 2014.
22. H. R. de Oliveira, A. T. Tavares, and B. F. Lóscio, "Feedback-based data set recommendation for building linked data applications," in *Proc. of the 8th ISWC*, pp. 49–55, ACM, 2012.
23. A. Nikolov and M. d'Aquin, "Identifying relevant sources for data linking using a semantic web index," in *WWW2011, LDOW*, 2011.
24. M. Mehdi, A. Iqbal, A. Hogan, A. Hasnain, Y. Khan, S. Decker, and R. Sahay, "Discovering domain-specific public SPARQL endpoints: a life-sciences use-case," in *Proc. of the 18th IDEAS 2014*, pp. 39–45.
25. L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze, "Identifying candidate datasets for data interlinking," in *Proc. of the 13th ICWE*, pp. 354–366, 2013.