Beyond Established Knowledge Graphs-Recommending Web Datasets for Data Linking

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

Beyond Established Knowledge Graphs-Recommending Web Datasets for Data Linking

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 explosive growth of the Web of Data in terms of size and complexity, identifying suitable datasets to be linked, has become a challenging problem for data publishers. To understand the nature of the content of specific datasets, we adopt the notion of dataset profiles, where datasets are characterized through a set of topic annotations. In this paper, we adopt a collaborative filtering-like recommendation approach, which exploits both existing dataset profiles, as well as traditional dataset connectivity measures, in order to link arbitrary, non-profiled datasets into a global dataset-topic-graph. Our experiments, applied to all available Linked Datasets in the Linked Open Data (LOD) cloud, show an average recall of up to 81%, which translates to an average reduction of the size of the original candidate dataset search space to up to 86%. An additional contribution of this work is the provision of benchmarks for dataset interlinking recommendation systems.

1 Introduction

The web of data, in particular Linked Open Data (LOD) [1], is growing constantly both in terms of size and impact. This growth introduces a wide variety and heterogeneity of Datasets with respect to represented resource types, currentness, coverage of topics and domains, size, used languages, coherence, accessibility [2] or general quality aspects [3]. The wide variety and heterogeneity of these dataset characteristics pose significant challenges for data consumers when attempting to find useful datasets without prior knowledge of available datasets. Dataset registries such as Datahub[1] or DataCite[2] aim at addressing this issue, for instance, by enabling users and data providers to annotate their datasets with some basic metadata, for instance, descriptive tags and access details. However, due to the reliance on human annotators, such metadata are often sparse and outdated [4]. This has contributed to the fact that, the majority of data consumption, linking and reuse focuses on established datasets and knowledge graphs such as DBpedia [5] or YAGO [6], while a long tail of datasets has hardly been reused and adopted.

[1] `http://datahub.io`
[2] `http://datacite.org`

For these reasons, dataset recommendation is becoming an increasingly important task to support challenges such as entity interlinking [7], entity retrieval or semantic search [8]. In line with [9], dataset recommendation is the problem of computing a rank score for each of a set of datasets $\mathcal{D}$ so that the rank score indicates the relatedness of a dataset from $\mathcal{D}$ to a given dataset, D_0. In turn, this allows to determine the likelihood of datasets in $\mathcal{D}$ to contain linking candidates for D_0.

While our approach is agnostic to the underlying data sharing principles, entity and data interlinking are of particular concern when considering Linked Open Data [1], not least because its essential principles and reliance on IRIs for identifying any term or entity facilitates Web-scale linking of data. Here, the current topology of the LOD cloud underlines the need for practical and efficient means to recommend suitable datasets, as only very few, well established knowledge graphs show a high amount of inlinks with DBpedia being the most obvious target, while a large amount of datasets is largely ignored, often due to a lack of understanding of their content and characteristics and consequently, the challenge to identify suitable linking candidates.

For the dataset recommendation problem, one has to consider both schema-level features, to take into account the overlap and complementarity of the actual schemas, as well as instance-level features, to consider the overlap and complementarity of described entities. Given the scale of available datasets, exhaustive comparisons of schemas and instances or some of their features are not feasible as an online process. Descriptive and reliable metadata, i.e. an index is required, which allows the efficient computation of suitable recommendations.

Some approaches exist, which obtain such an index through topic modeling approaches. For instance, [4] generates a weighted bipartite graph, where datasets and topics represent the nodes, related through weighted edges, indicating the relevance of a topic for a specific dataset. However, while computation of such *topic profiles* is costly, it is usually applied to a subset of existing datasets only, where any new or so far unannotated datasets require the pre-computation of a dedicated topic profile.

In our work, we provide a recommendation method which not only takes into account the direct relatedness of datasets as emerging from the topic-dataset-graph produced through the profiling in [4], but instead, we adopt established *collaborative filtering* practices by considering the topic relationships emerging from the global topic-dataset-graph to derive specific dataset recommendations. We exploit dataset connectivity measures to relate non-profiled datasets to datasets in the dataset-topic-graph, enabling us to consider arbitrary datasets as part of our recommendations. This approach on the one hand significantly increases the recall of our recommendations, but at the same time improves recommendations through considering dataset connectivity as another relatedness indicator. The intuition is that this leads to more robust and less error-prone recommendations, since the consideration of global topic connectivity provides reliable connectivity indicators even in cases where the underlying topic profiles might be noisy. Our

assumption is that even poor or incorrect topic annotations will serve as reliable relatedness indicator when shared among datasets.

While we adopt the topic profile graph in [4] for our experiments, we would like to emphasize that our approach is agnostic to the underlying topic index. Topic profiles which are obtained by annotating samples of instances as in the chosen method, are shown to reflect both, instance-level as well as schema-level characteristics of a specific dataset. Even though topics are derived from instances, resources of particular types show characteristic topic distributions, which significantly differ across different types [10].

In our experiments, we apply our approach to the LOD cloud as one scenario and use case, where dataset recommendation is of particular relevance. Our experiments show superior performance compared to three simple baselines, namely based on shared key-words, shared topics, and shared common links. In a series of experiments, we demonstrate the performance of our technique compared to the current version of the LOD as an evaluation data, achieving a reduction of the original (LOD) search space of up to 86% on average.

We proceed to present the theoretical grounds of our technique in Section 2, which contains *two* of the contributions of the paper – an efficient approach of propagating dataset profiles over the LOD cloud by starting off with a small set of profiled datasets and a dataset recommendation technique based on topic-profiles. Section 3 defines the evaluation framework that has been established and reports on our experimental results, providing a comparison to a set of baseline recommendation approaches, made available, as a *third* contribution of this work, to the community as a benchmark. Related approaches are presented and discussed in Section 4 before we conclude in Section 5.

2 Dataset Recommendation Framework

The current section introduces a novel approach to dataset recommendation based on dataset profiles with an aimed application in the entity interlinking process. In the current setting, the datasets profiles are generated by a topic modeling paradigm, which is briefly introduced in the following subsection together with some notation and basic definitions. A computationally efficient approach of propagating existing profiles towards new arbitrary datasets is presented in subsection 2.2. Our recommendation technique is given in detail in subsection 2.3. Fig. 2 provides an overview of the main steps of the approach that will be discussed in the sequel.

2.1 Preliminaries

We start by introducing notation and definitions. Let $T_1, ..., T_N$ be a number of topics from a set of topics $\mathcal{T}$ and let $\mathcal{D} = \{D_1, ..., D_M\}$ be a set of datasets.

Dataset Topic Profile Topic modeling algorithms such as Latent Dirichlet allocation [11] are used to discover a set of topics from a large collection of

documents, where a topic is a distribution over terms that is biased around those associated under a single theme. Topic modeling approaches have been applied to tasks such as corpus exploration, document classification, and information retrieval. Here, we will look into a novel application of this group of approaches, exploiting the topic structure in order to define and construct dataset profiles for dataset recommendation.

As a result of the topic modeling process, a bipartite—*profile*—graph is built, providing a relation between a document and a topic. Documents in our setting are the datasets to be considered, therefore the profile graph is induced by the relation between a dataset, D_i, and a topic, T_k, expressed by a weight, $w_{ik} \in [0,1]$, for all $i = 1, ..., M$ and $k = 1, ..., N$. Formally, a profile graph is defined as follows.

Definition 1 (Dataset Topic Profile Graph). *A dataset topic profile graph is a weighted directed bipartite graph $\mathcal{P} = (\mathcal{S}, \mathcal{E}, \Delta)$, where $\mathcal{S} = \mathcal{D} \cup \mathcal{T}$, $\mathcal{E}$ is a set of edges of the form $e_{ik} = (D_i, T_k)$ such that $D_i \in \mathcal{D}$ and $T_k \in \mathcal{T}$ and*

$$\begin{aligned} \Delta \colon \mathcal{E} &\to [0,1] \\ e_{ik} &\mapsto w_{ik} \end{aligned}$$

is a function assigning weights to the edges in $\mathcal{E}$.

The bipartite property of $\mathcal{P}$ allows to represent a given dataset by a set of topics—its *profile*. For the purposes of this study, it is worth noting that, inversely, a topic can be represented by a set of weighted datasets—what we will call the *signature* of a topic (see Figure 1). We will denote by Profile(D_i) the function returning the topic profile of D_i, i.e., the set of topics together with their weights with respect to D_i. Inversely, we will denote by $\mathcal{D}_{T_k}$ the set of datasets together with their weights with respect to a topic T_k, derived again from the graph $\mathcal{P}$.

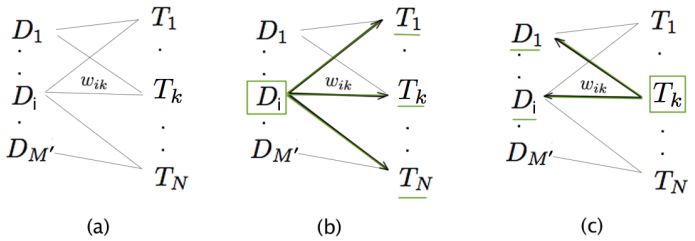

Fig. 1. (a) An example of a bipartite profile graph with topics and datasets linked by weighted edges. (b) Representing a dataset, D_i, as a set of topics. (c) Representing a topic, T_k, as a set of datasets.

Datasets Connectivity The connectivity behavior of datasets is a central concept within the proposed recommendation framework. We consider the following definition of a measure of the strength of dataset connectedness.

Definition 2 (Dataset inter-connectivity measure). *Let $D_i, D_j \in \mathcal{D}$ be two datasets. We define a measure of their common degree of connectivity as follows.*

$$\mathcal{C}(D_i, D_j) = \frac{shared(D_i, D_j) \times [total(D_i) + total(D_j)]}{2 \times total(D_i) \times total(D_j)} \tag{1}$$

where $shared(.,.)$ returns the number of links between two datasets and $total(D_i)$ returns the total number of links between D_i and any other dataset in $\mathcal{D}$.

Note that (1) is the symmetric version of the measure of connectivity of D_i to D_j given by

$$\mathcal{C}'(D_i, D_j) = \frac{shared(D_i, D_j)}{total(D_i)}.$$

Explicitly, (1) is obtained by taking the mean

$$\frac{\mathcal{C}'(D_i, D_j) + \mathcal{C}'(D_j, D_i)}{2} = \mathcal{C}(D_i, D_j).$$

The measure $\mathcal{C}$ is in the interval $[0, 1]$ and has the advantage of considering the relative connectivity between datasets instead of simply looking at the number of links. In our experimental setting, $shared(D_i, D_j)$ is taken as the sum of the links between two datasets in both directions: $D_i \rightarrow D_j$ and $D_j \rightarrow D_i$, resulting in the number of incoming and outgoing links between the datasets. A specific version of the measure C can be defined by taking only certain types of links (or predicates) in consideration (in our application scenario, we have considered LOD datasets, therefore an example of a specific predicate can be `owl:sameAs`).

In a more general manner, it is possible to use any dataset connectivity measure of our choice. The measure given above is one that worked well in our experiments (see Section 3). In addition, one can define in a broader sense a measure of dataset relatedness incorporating semantic elements such as vocabulary and keywords overlap. Dataset complementarity can be of interest in certain scenarios, as well. However, in the current study we have focused on connectivity only, leaving the other possibilities out for future work.

2.2 The Preprocessing/Learning Step

In many cases the number of indexed elements (e.g., datasets in our case) is much lower than the entire number of elements of interest. In that respect, it is interesting to consider a procedure that allows inexpensively to include novel elements in the index. As a preprocessing step of our recommendation workflow, we adopt a learning approach that consists of assigning topics to datasets by linking them into the dataset-topic-graph after computing their connectivity

with already indexed (profiled) datasets. This step is useful in order to include in the recommendation pipeline datasets that have not been initially indexed, in an inexpensive manner, keeping in mind that the indexing process can be often quite costly and time-consuming.

Let $\mathcal{P}$ be a topic profile graph and let $D_j \in \mathcal{D}$ be a random dataset, which is not necessarily included in the original topic profile graph $\mathcal{P}$. We assign topic weights to D_j considering its degree of connectivity with respect to datasets from the topic profile graph by using the following measure of relatedness between linked datasets and topics (see Fig. 2, steps 1 and 2).

Definition 3 (Connectivity-based dataset and topic relatedness measure). *Let $D_j \in \mathcal{D}$ and $T_k \in \mathcal{T}$. We define the following dataset and topic relatedness measure.*

$$\sigma(D_j, T_k) = \max_{D_i \in \mathcal{D}} \mathcal{C}(D_i, D_j) * w_{ik}. \tag{2}$$

Recall that w_{ik} is the weight of the topic T_k with respect to D_i as given in Def. 1, taking a zero value in case T_k is not in Profile(D_i). $\mathcal{C}(D_i, D_j)$ is the connectivity measure between two datasets, as defined in (1). The dataset and topic relatedness measure σ is a way to measure the datasets connectivity behavior using their profiles. We will use the notation $\sigma_{jk} = \sigma(D_j, T_k)$ as a shortcut. Note that σ is in the $[0, 1]$ interval.

This new weighting scheme allows to propagate inexpensively the profile of D_i to datasets that are connected to it. Hence, a new graph is created between target datasets and source datasets topics. Precisely, a topic $T_k \in$ Profile(D_i) will be assigned to a dataset D_j that has a non-zero value of $\mathcal{C}(D_i, D_j)$. The weight of this novel topic-dataset relation is now based on the connectivity order of D_j with respect to D_i, scaled by the weight w_{ik} of T_k with respect to D_i. In that sense, w_{ik} plays a penalization role: the novel weight σ_{jk} of T_k with respect to D_j is penalized by the weight of T_k in the original topic graph, i.e., datasets with high degree of connectivity to D_i will get relatively low weights with respect to a topic, if that topic has a relatively low weight with respect to D_i. We consider the maximum value over all datasets in $\mathcal{D}$, the set of the originally profiled datasets. In this way, we avoid ambiguity when a non-indexed dataset D_j is connected to a single topic T_k via multiple already indexed datasets, assuring that the highest value of relation between T_k and D_j is preserved. Thus, the choice of a topic to be assigned to a dataset is not influenced, only its weight is, and no connectivity information is lost.

The topic-dataset relatedness measure (3) allows to construct a novel profile graph by computing σ_{jk} for all possible values of j and k ($j = 1, ..., M$ and $k = 1, ..., N$). The novel graph, that we call *the Linked Dataset Topic Profile Graph (LDPG)*, includes new datasets and the original topics as its nodes and is defined as follows (see Fig. 2, step 2).

Definition 4 (Linked Dataset Topic Profiles Graph (LDPG)). *The LDPG is a weighted directed bipartite graph $\mathcal{P}_l = (\mathcal{S}_l, \mathcal{E}_l, \Delta_l)$, where $\mathcal{S}_l = \mathcal{D} \cup \mathcal{T}$, $\mathcal{E}_l$ is a*

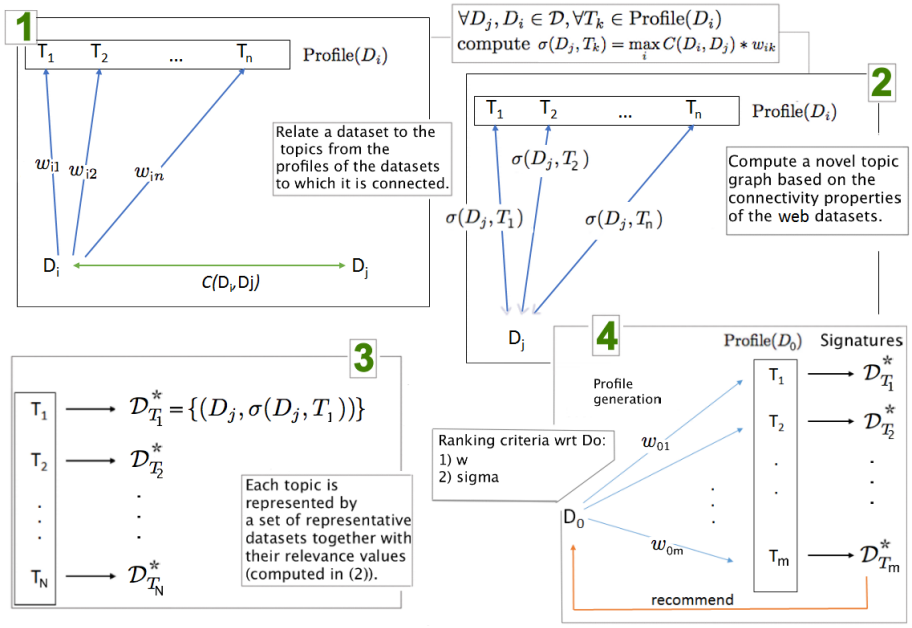

Fig. 2. The four main steps of the profile-based dataset recommendation framework.

set of edges of the form $e'_{jk} = (D_j, T_k)$ *such that* $D_j \in \mathcal{D}$ *and* $T_k \in \mathcal{T}$ *and*

$$\Delta_l : \mathcal{E}_l \to [0,1]$$
$$e'_{jk} \mapsto \sigma_{jk}$$

is a function assigning weights to the edges in $\mathcal{E}_l$.

As this was the case within the original profiling scheme, the inherently bipartite nature of the graph $\mathcal{P}_l$ allows for a two-fold interpretation — either a dataset is modeled as a set of topics (a dataset's *profile*), or, inversely, a topic is modeled as a set of datasets assigned to it (a topic's *signature*). Therefore, it is easy to define a set of **significant** datasets with respect to a given topic, by thresholding on their weights in the Linked profiles graph with respect to the topic of interest. Note again that for the purposes of the recommendation task, we will be interested in keeping the weights of every dataset in the resulting topic representations and thus model every topic by a set of *(dataset, weight)* couples.

Definition 5 (Dataset significance for a topic. Topic signature). *A dataset* $D_j \in \mathcal{D}$ *is* **significant** *with respect to a topic* $T_k \in \mathcal{T}$ *if its weight in the LDPG* $\sigma_{jk} = \sigma(D_j, T_k)$ *is greater than a given value* $\theta \in (0,1)$.

A topic T_k *is modeled by the set of its significant datasets together with their respective weights, given as*

$$\mathcal{D}^*_{T_k} = \{(D_j, \sigma_{jk}) | \sigma_{jk} > \theta\}_{j=1,\ldots,M}, \tag{3}$$

for $k = 1, ..., N$. *We will call* $\mathcal{D}^*_{T_k}$ *the* **signature** *of the topic* T_k.

With this definition, the profile of a given dataset, Profile(D_i), is modeled as a number of sets of significant datasets – one per topic in Profile(D_i) coupled with their weights with respect to each topic (see Fig. 2, step 3), or otherwise – a set topic signatures.

For sake of generality, we draw the readers attention to the fact that the learning approach resulting in index extension applies to any dataset profile definition that one might like to consider and not exclusively to the one based on the topic modeling paradigm.

2.3 Profile-Based Dataset Ranking

Let D_0 be a new dataset to be linked. The aim of the recommendation task is to provide the user with an ordered list of datasets, potential candidates for interlinking with D_0, which narrows down considerably the original search space i.e., the web of data. Thus the dataset recommendation can be seen as the problem of computing a rank score for each $D_j \in \mathcal{D}$ that indicates the likelihood of D_j to be relevant to a dataset D_0. In the context of using topic-based dataset profiles for linking recommendation, we restate the problem in the following manner.

> *For a given non-linked dataset* D_0, *profile-based dataset recommendation is the problem of computing a rank score* r_{0j} *for each* $D_j \in \mathcal{D}$ *based on topic overlap between* D_j *and* D_0, *so that* r_{0j} *indicates the relevance of* D_j *to* D_0 *for the entity linking task.*

We start by generating the topic profile of D_0, Profile(D_0) = $\{(T_1, w_{01}), ..., (T_m, w_{0m})\}$. Then, we extract from the result of the learning step the set of *target* datasets for each topic in Profile(D_0) together with their corresponding relevance values σ, namely the set of m topic signatures $\{\mathcal{D}^*_{T_k}\}_{k=1}^m$. These datasets constitute the pool, from which we will recommend interlinking candidates to D_0. We will use n to denote their number, that is $n = \sum_{j=1}^{m} |\mathcal{D}^*_{T_j}|$, or the sum of the numbers of datasets in each topic signature. The aim is to serve the user with the most highly ranked datasets from that pool. There are two ranking criteria to consider: the weight w of each topic in Profile(D_0) and the weight σ of each dataset in each of the topic signatures in $\{\mathcal{D}^*_{T_k}\}_{k=1}^m$ (step 4 in Fig. 2). Since the ranking score in our setting depends on topic overlap, we define the interlinking relevance of a dataset D_j with respect to D_0 in the following manner.

Definition 6 (Dataset interlinking relevance). *For all* $j = 1, ..., n$, *the relevance of a dataset* $D_j \in \mathcal{D}$ *to a dataset* D_0 *via the topic* T_k *is given by*

$$r_j^0 = w_{0k} * \sigma_{jk}, \tag{4}$$

with $k = 1, ..., m$.

Note that j covers the total number of datasets in the set of m topic signatures, therefore the relevance value depends on j only (i.e., a single relevance value per dataset from the pool of candidates). Similarly to the definition of σ in Def. 6, w has a penalization function, decreasing the ranks of datasets that have high values of their σ weights, but are found in topic signatures of a low relevance to D_0 (expressed by a low value of w).

It is easy to define a mapping $f : \mathcal{R} \to \mathbb{N}$ from a space of interlinking relevance values $\mathcal{R}$ to the natural numbers such that $f(r^0_{j_1}) > f(r^0_{j_2}) \iff r^0_{j_1} \leq r^0_{j_2}$, for any $j_1, j_2 \in [1, n]$ and $1 = \max_j f(r^0_j)$. With this definition, since there is a relevance value r^0_j per dataset $D_j \in \mathcal{D}$, $f(r^0_j)$ returns the rank of the dataset D_j with respect to D_0. The results of the recommendation process are given in a descending order with respect to these ranks.

3 Experiments and Results

In this section, we start by a discussion on the evaluation setting then 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.* [12] provide a large review of recommender system evaluation techniques and cite three common types of experiments: (i) offline setting, where recommendation approaches are compared without user interaction, (ii) user studies, where a small group of subjects experiment with the system and report on the experience, and (iii) online experiments, where real user populations interact with the system.

In our approach, we assume that the dataset connectivity behavior when data were collected (i.e., steps 1, 2 and 3 in Fig. 2) is similar enough to the profile connectivity behavior when the recommender system is deployed (i.e., step 4 in Fig. 2), so that we can make reliable decisions based on an *offline evaluation*. The offline experiment is performed by using pre-collected data as evaluation data (ED). Using these data, we can simulate the profiles connectivity behavior that impacts the recommendation results.

The most straightforward, although not unproblematic (see the discussion that follows below) choice of ED for the entity linking recommendation task is the existing link topology of the current version of links between web datasets. Since this evaluation data are the only available data that we have for both training (our preprocessing steps 1, 2 and 3 in Fig. 2) and testing (the actual recommendation in step 4 of Fig. 2), we opted for a *5-fold cross-validation* [13] to evaluate the effectiveness of the recommendation system. In 5-fold cross-validation, the ED was randomly split into two subsets: the first one, containing random 80% of the linked datasets in the ED, was used as training set while the second one, containing the remaining linked datasets (i.e., random 20% of

the ED), was retained as the validation data for tests (i.e., the test set). We repeated these experiments five times changing at each time the 20% representing the test set in order to cover 100% of the whole data space. The evaluation is based on the capacity of our system to *reconstruct* the links from the ED in the recommendation process.

The most common measures of the efficiency of a recommendation system are Precision, Recall and F1-Score, formalized as functions of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) as follows.

$$Pr = \frac{TP}{TP + FP}; \quad Re = \frac{TP}{TP + FN}; \quad F1 = \frac{2TP}{2TP + FN + FP}. \tag{5}$$

In addition, [12] present a measure of the false positive overestimation, particularly important in our offline evaluation case:

$$FalsePositiveRate = \frac{FP}{FP + TN}. \tag{6}$$

A small value of the false positive rate means that every time you call a positive, you have a high probability of being right. Conversely, a high value of the false positive rate means that every time you call a positive, you have a high probability of being wrong.

3.2 Experimental Setup

Since our recommendation is based on **the connectivity** of the $\mathcal{D}$ graph as well as **the topic profiles** of indexed datasets, the Linked Open Data (LOD) [1] is clearly our best use case to experiment our recommendation. As a topic modeling approach, we adopt the dataset profiles index provided by [4] since it generates accurate profiles and outperforms established topic modeling approaches such as ones based on the well known Latent Dirichlet Allocation [3]. Using the topic profiles approach from [4] it is easy to produce a weighted bipartite graph as described in 1, where datasets and topics represent the nodes, related through weighted edges, indicating the relevance of a topic for a specific dataset. A *dataset profile* is represented through *topics*, which in this case are DBpedia categories[4] derived through a processing pipeline[5] analyzing representative resource samples from specific datasets. For example, the $\mathcal{T}$ profile of the `Semantic Web Dog Food Corpus`[6] dataset includes the following DBpedia categories: *Data management, Semantic Web, Information retrieval, World Wide Web Consortium standards, etc.*

In the following, we distinguish between the set $\mathcal{D}$ of datasets in the entire LOD and the datasets indexed by the profiling approach [4], denoted by $\mathcal{D}'$. Explicitly, we consider:

[3] http://mallet.cs.umass.edu/
[4] `dbpedia.org/page/Category`
[5] `http://data-observatory.org/lod-profiles/profiling.htm`
[6] `http://data.semanticweb.org/`

– $\mathcal{D}$: All datasets in the LOD cloud group on the Data Hub[7], which will be considered as target datasets (to be recommended) in the testing set, $|\mathcal{D}| = 258$.
– $\mathcal{D}'$: All the datasets indexed by the topics profiles graph[8], which will be considered as source datasets (to be linked) in the testing set, $|\mathcal{D}'| = 76$ and $|\mathcal{D}'| \subset |\mathcal{D}|$.

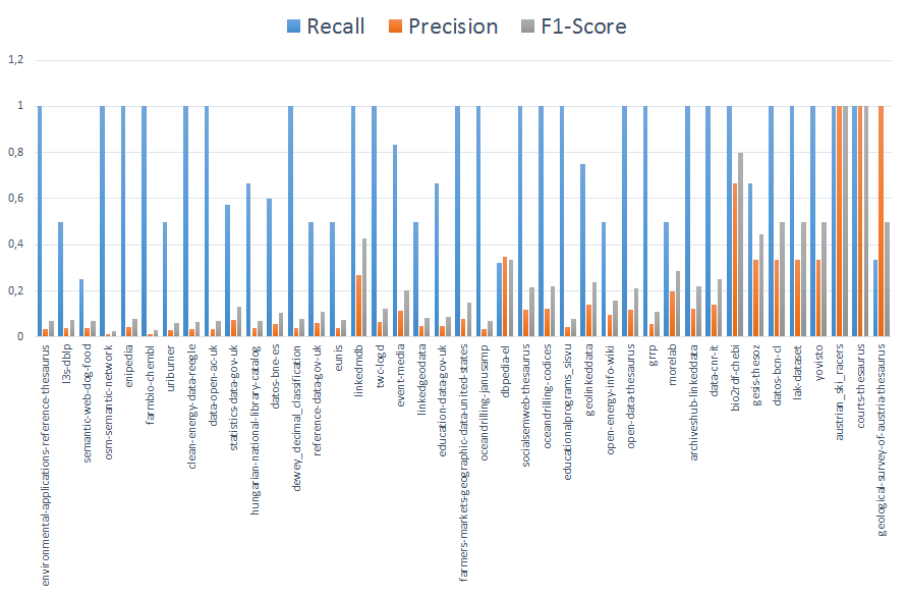

Fig. 3. Recall/Precision/F1-Score over all recommendation lists for all source datasets in $\mathcal{D}'$ and all target datasets in $\mathcal{D}$.

We trained our system as described in steps 1, 2 and 3 in Fig. 2. We started by extracting the topic profiles graph from the available endpoint of Data Observatory[9]. Then we extracted VoID descriptions of all LOD datasets, using the *datahub2void* tool[10]. The constituted **evaluation data (ED)** corresponds to the outgoing and incoming links extracted from the generated VoID file (it is made available on `http://www.lirmm.fr/benellefi/void.ttl`).

Note that in the training set we used the actual values of `VoID:triples` (see Section 1) to compute dataset connectivity, while in the test set we considered binary values (two datasets from the evaluation data are either linked or not).

[7] `http://datahub.io/group/lodcloud`
[8] `http://data-observatory.org/lod-profiles/profile-explorer/`
[9] `http://data-observatory.org/lod-profiles/sparql`
[10] `https://github.com/lod-cloud/datahub2void`

For example, $shared(tip, linkedgeodata) = 6$, so in the training set we considered 6 as the number of links in Eq. (1), while in the test set we only consider the information that *tip* is connected to *linkedgeodata* and vice versa. Training is performed only once.

3.3 Results and Analysis

We ran our recommendation workflow as described in step 4 in Fig. 2. Using 5-fold cross-validation, for each dataset in $\mathcal{D}'$, we recommended an ordered list of datasets from $\mathcal{D}$. The results are given in Fig.3.

The results show a high average recall of up to 81%. Note that the recommendation results for 59% of the source datasets have a recall of 100% and two of them have an F1-score of 100%. As mentioned in Section 3.2, we considered only the binary information of the existence of a link in the LOD as evaluation data in the testing set. This simplification has been adopted due to the difficulty of retrieving all actual links in the LOD graph (implying the application of heavy instance matching or data linking algorithms on a very large scale). Certainly, the explicit currently existing links are only a useful measure for recall, but not for precision. In our experiments, we measured an average precision of 19%. We explain that by the fact that the amount of explicitly declared links in the LOD cloud as ED is certain but far from being complete to be considered as ground truth. Subsequently, 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. For this reason, based on our evaluation data, a large amount of false positives occur, which in reality are likely to be relevant recommendations. In order to rate this error, we calculated the false positive rate over all results, shown in the Fig. 4. The small values of this rate indicate that every time you call a positive, you have a probability of being right, which provide support to our hypothesis with an average FP-Rate of 13%.

To further illustrate the effect of false positives overestimation, we included in the ED new dataset links based on the shared keywords of the datasets. Precisely, if two datasets share more then 80% of their VoID tags, they are considered as linked, and are added to the ED. For example, *linkedgeodata* is connected to 4 datasets in the main ED: *osm-semantic-network, dbpedia, tip et dbpedia-el*. However, we found that *linkedgeodata* shared more than 80% of its tags with *fu-berlin-eurostat* and *twarql*[11]. By adding both links to the original ED, we noted a gain in precision of 5% for the *linkedgeodata* dataset with no impact on recall. Thus, we believe that our system can perform much better on more complete ED.

The current version of the topic dataset profile graph from [4] contains 76 datasets and 185 392 topics. Working with this already annotated subset of existing datasets is not sufficient and limited the scope of our recommendations

[11] Example: *linkedgeodata* has 11 tags and *twarql* has 9 tags. We considered as connected since they shared 8 tags which is higher than the 80% of the average amount, i.e., $8 < (0.8 * (11 + 9)/2)$.

significantly. In addition, the number of the profiled datasets, compared to the number of topics is very small, which in turn appeared to be problematic in the recommendation process due to the high degree of topic diversity leading to a lack of discriminability. One way of approaching this problem would be to index all LOD datasets by applying the original profiling algorithm [4]. However, given the complexity of this processing pipeline—consisting of resource sampling, analysis, entity and topic extraction for a large amount of resources—it is not efficient enough, specifically given the constant evolution of Web data, calling for frequent re-annotation of datasets. Also we note that this profiles propagation technique can be of interest for any dataset profile-based application, providing an inexpensive way of computing a profile of an arbitrary dataset. As one of the original contributions of this paper, the preprocessing step in our recommendation pipeline can be seen as an efficient method for automatic expansion of the initial profiles index over the entire linked open data space based on dataset connectivity measures.

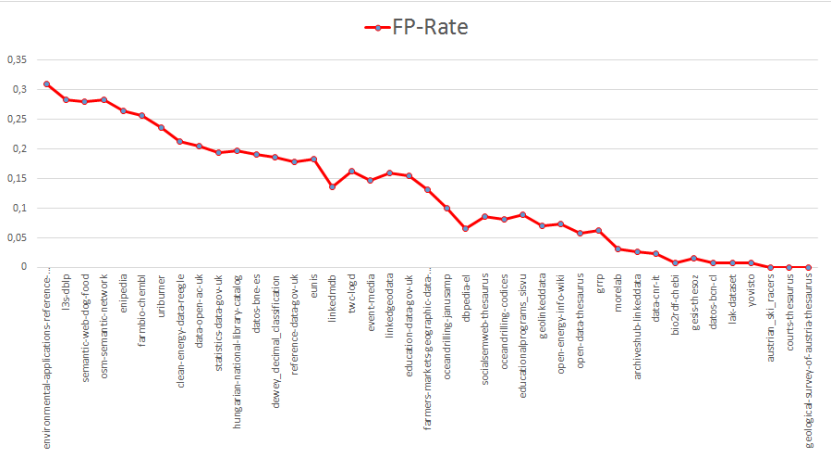

Fig. 4. False Positive Rates over all recommendation lists over all $\mathcal{D}'$ datasets.

The main goal of a recommender system is to reduce the cost of identifying candidate datasets for the interlinking task. Some systems may provide recommendations with high quality in terms of both precision and recall, but only over a small portion of datasets (as is the case in [14]). We obtain high recall values for the majority of datasets over *the entire set of LOD datasets* with a price to pay of having relatively low precision. Here, low precision/high recall systems still offer significant contributions by narrowing the size of the search space. Therefore, we highlight the efficiency of our system in reducing the search space size. Fig. 5 depicts the reduction of the original search space size (258 datasets) in percentage over all source datasets to be linked. The average space size reduction is of up to 86%.

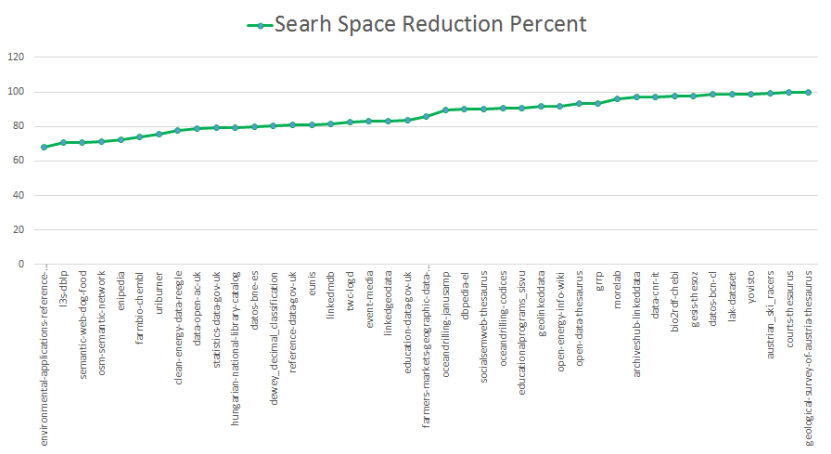

Fig. 5. Search space reduction in percent over all recommended sets and over all $\mathcal{D}'$ datasets.

As mentioned previously, our system can cover 100% of the available linked datasets, since the topics-datasets profiling approach [4] as well as our profile expansion approach presented in Section 2.2 are able to profile any arbitrary dataset. Our system is also capable of dealing with the well-known cold-start problem (handling correctly newly published and unlinked datasets), since not linked datasets are handled by using the indexing technique in [4], which does not rely on dataset connectivity. Based on the learning step, our system is able to recommend a sorted list of candidate datasets to these not linked datasets.

3.4 Baselines and Comparison

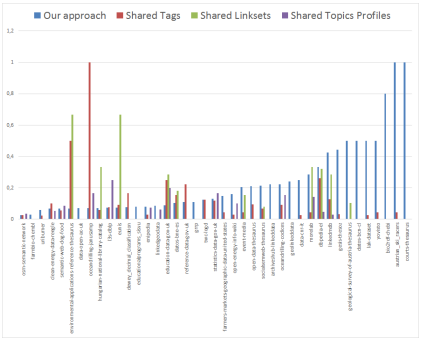

Fig. 6. F1-Score values of our approach versus the baselines overall $\mathcal{D}'$ datasets

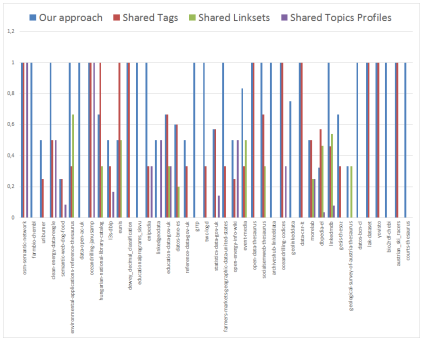

Fig. 7. Recall values of our approach versus the baselines overall $\mathcal{D}'$ datasets

To the best of our knowledge, there does not exist a common benchmark for dataset interlinking recommendation. One of the contributions of this paper is the provision of three simple baseline approaches for this problem. Given two datasets, D_0 and D_j, we define the following baseline recommendation methods.

Shared Keywords Recommendation: if D_0 and D_j share N_{tags} of VoID:Tags extracted from `http://www.lirmm.fr/benellefi/void.ttl` with $N_{tags} > 0$, then we recommend (D_j, N_{tags}) to D_i, where N_{tags} acts as a rank score.

Shared Links Recommendation: if D_0 and D_j have N_{links} connected datasets in common from `http://www.lirmm.fr/benellefi/void.ttl` with $N_{linksets} > 0$, then we recommend (D_j, N_{links}) to D_0, where N_{links} acts as a rank score.

Shared Topics Recommendation: if D_0 and D_j share N_{topics} topics extracted from `http://data-observatory.org/lod-profiles/sparql` with $N_{topics} > 0$, then we recommend (D_j, N_{topics}) to D_0, where N_{topics} acts as a rank score.

The recommendation results for all LOD datasets (D_0 covering D) of the three baseline approaches are made available on `http://www.lirmm.fr/benellefi/Baselines.rar`.

	Our approach	Shared Keywords	Shared linksets	Shared Topics Profiles
AVG Precision	**19%**	9%	9%	3%
AVG Recall	**81%**	47%	11%	13%
AVG F1-Score	**24%**	10%	8%	4%

Table 1. Average precision, recall and F1-score of our system versus the baselines over all $\mathcal{D}'$ datasets based on the ED.

Fig. 6 and Fig. 7, respectively, depict detailed comparisons of the F1-Score and the Recall values between our approach and the baselines over all $\mathcal{D}'$ datasets taken as source datasets. From these figures, it can be seen that our method largely outperforms the baseline approaches, which even fail to provide any results at all for some datasets. The baseline approaches have produced better results than our system in a limited number of cases, especially for source and target datasets having the same publisher. For example, the shared keywords baseline generated an F-Score of 100% on *oceandrilling-janusamp*, which is connected to * *oceandrilling-codices*, due to the fact that these two datasets are tagged by the same provenance (`data.oceandrilling.org`).

Table 1 compares the performance of our approach to the three baseline methods in terms of average precision, recall and F1-score.

As a general conclusion, these observations indicate that the collaborative filtering-like recommendation approach, which exploits both existing dataset profiles as well as traditional dataset connectivity measures, shows high performance on identifying candidate datasets for the interlinking task.We make all of the ranking results of our recommendation approach available to the community on `http://www.lirmm.fr/benellefi/results.csv`.

4 Related Work

With respect to finding relevant datasets on the Web, we cite briefly several studies on discovering relevant datasets for query answering have been proposed. Based on well-known data mining strategies, the works in [15] and [16] present techniques to find relevant datasets, which offer contextual information corresponding to the user queries. A used feedback-based approach to incrementally identify new datasets for domain-specific linked data applications is proposed in [17]. 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 directly relevant to our work.

Nikolov *et al.* [18] propose a keyword-based search approach to identify candidate sources for data linking. The approach consists of two steps: (i) searching for potentially relevant entities in other datasets using as keywords 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.

Leme *et al.* [14] 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 (*cf.* the cold-start problem discussed previously). 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.

Mehdi *et al.* [19] 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 for exact literal matches between the introduced keywords and the target datasets, i.e., for each term supplied to the algorithm, the system runs a matching with 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 particularly useful in the linking step.

In contrast to the approaches described above, our method is the first to be based on topic overlap and collaborative filtering. To the best of our knowledge, the current paper is also the first study to provide simple baseline recommendation techniques that can serve as a benchmark (*cf.* Section 3.4).

In comparison with all the work discussed above, our approach has the potential to overcome a series of complexity related problems. Precisely, considering

the complexity to generate the matching in [18], to produce the set of domain-specific keywords as input in [19] 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 already generated topic profiles (or inexpensively propagated profiles).

Since the majority of used datasets in the papers discussed above were not yet indexed by the topic profiles graph, a direct comparison of the performance of the different recommendation methods to our system seems unfair. In addition, none of the authors have shared data, except for [9]. However, the evaluation of this approach in its first version [14] concludes on one dataset only. In its second version [9], while they provide the data used in their experiments, the authors do not give details on the resulting rank scores.

5 Conclusion and Future Work

We have presented an interlinking candidate dataset recommendation approach, based on the connectivity behavior learning of topic-profiles datasets. We demonstrate that our technique allows to reduce considerably the original search space and that it outperforms the results obtained by three baseline recommendation approaches, developed for the purposes of this study and made available to the community. Since dataset topic profiling is a key component in the recommendation pipeline, we show a simple way of propagating dataset profiles to the entire set of linked open datasets, starting off with a limited number of profiled datasets.

In the future, we plan to improve the evaluation framework by developing a more reliable and complete ground truth for dataset recommendation, possibly by using crowdsourcing techniques, in order to deal with the false positives overestimation problem. Our method could potentially benefit from combining it with machine learning techniques. We plan to conduct a thorough evaluation of the efficiency of our profiles propagation technique for the dataset recommendation task.

Acknowledgements

This research has been partially funded under the Datalyse project[12] and by the European Commission as part of the DURAARK project, FP7 Grant Agreement No. 600908.

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.

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

2. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbusshe, "Sparql web-querying infrastructure: Ready for action?," in *Proc. of the 12th ISWC, Sydney, Australia*, 2013.
3. C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," in *Proc. of the 9th ESWC - The Semantic Web: Research and Applications*, pp. 87–102, 2012.
4. 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 *Proc. of the 11th ESWC*, Springer, 2014.
5. 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.
6. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proc. of WWW*, pp. 697–706, 2007.
7. 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.
8. 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.
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. D. Taibi, S. Dietze, B. hu, and G. Fulantelli, "Exploring type-specific topic profiles of datasets: a demo for educational linked data," in *Proc.of the ISWC Posters & Demonstrations Track a track, Riva del Garda, Italy.*, pp. 353–356, 2014.
11. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
12. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, *Recommender systems handbook*, vol. 1. Springer, 2011.
13. S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991.
14. 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, Aalborg, Denmark*, pp. 354–366, 2013.
15. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Discovering related data sources in data-portals," in *Proc. of the 1st IWSS*, 2013.
16. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Entity-based data source contextualization for searching the web of data," in *Proc. of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th ESWC Satellite Events, Crete, Greece.*, pp. 25–41, 2014.
17. 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.
18. A. Nikolov and M. d'Aquin, "Identifying relevant sources for data linking using a semantic web index," in *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India*, 2011.
19. 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, Porto, Portugal*, pp. 39–45.