Recommending spatial classes for entity interlinking in the Web of Data

Vasilis Kopsachilis[1] [0000-0003-3824-3932]

[1] University of Aegean, Mytilene, Greece
vkopsachilis@geo.aegean.gr

Abstract. Recent advances in the web informatics domain bring closer the realization of Web of Data, a global interconnected data space where richer entity descriptions are easily retrievable and reusable. A key Web of Data component is the establishment of links between related entities. Link Discovery tools can be utilized for the (semi) automatic identification and linkage of related entities between a pair of entity sets. However, they require the manual examination and selection of Web of Data datasets (or sub parts of them) that will be used for link establishment. This research focuses on proposing automated methods, which search in Web of Data datasets and recommend pairs of classes that may contain related entities and thus can be used as input in Link Discovery tools. We approach the problem from a geographical perspective by exploiting the spatial information of classes i.e. the location of their instances. We intuitively believe that classes that present similar spatial distribution is likely to contain related entities. To achieve scalability at web scale, we study and implement spatial summarization methods that capture the spatial distribution of each class. To identify relevant classes, we investigate and propose techniques that act on the summaries to compute their similarity. We a) evaluate two aspects of our methodology, namely the ability of identifying relevant classes effectively and performing at web scale efficiently and b) compare our approach with other state of the art dataset recommendation for interlinking approaches.

Keywords: Web of Data, Spatial Data, Dataset Recommendation, Entity Interlinking

1 Introduction

Over the last years many data providers have been publishing their data on the web according to the Linked Data principles [1] weaving the Web of Data, a global entity-centric data space where entities across the web are more discoverable and easier reusable [2]. A fundamental prerequisite for the realization of the Web of Data is the establishment of links between, dispersed across different datasets, entities for which a kind of relation is hold (e.g. they refer to the same real world object). Towards the goal of link establishment, Linked Data best practices suggest data providers to apply to their data Link Discovery methodologies, implemented by tools such as SILK [3] or LIMES [4]. Link Discovery refers to the process of identifying and interlinking

related entities between two (or more) given datasets (or more abstractly entity sets) [5]. A preprocessing step in the Link Discovery workflow requires data providers to provide as input a pair of entity sets that will be used for link establishment. Therefore, data providers should have prior knowledge of the available in the Web of Data datasets and their sub parts that may contain related entities. This PhD focuses on this preprocessing step of Link Discovery workflow, i.e. the identification of Web of Data datasets and sub parts of them that may contain related entities for interlinking. The identified entity sets can be then used as input in Link Discovery tools.

Several works give insights about the Web of Data size and connectivity status. The last version of LOD cloud diagram [6], created in 2017, was including 1.163 datasets. LODStats [7], in order to generate Web of Data statistics, parsed about 3.000 datasets containing in total approximately 50 million entities. A deeper Web of Data analysis [8] reveals that 44% of datasets do not contains links to other datasets. Furthermore, only a small number of datasets is highly linked while the majority is only sparsely linked [9]. Data providers tend to link their datasets with well-known datasets (such as DBpedia or GeoNames) and ignore less well known datasets which may also contain related entities for interlinking [10]. As [11] points linkage with popular datasets is favored because of two main reasons: (i) the difficulty in finding related datasets; and (ii) the strenuous task of discovering instance mappings between different datasets. Data providers can look up for relevant datasets by examining the LOD cloud diagram, which provides an overview of the datasets domain and connectivity, or by querying dataset catalogs, such as datahub.io,[1] which preserve user submitted datasets metadata. However, since Web of Data is continuously expanding (LOD cloud reports a 294% increase in the number of the LOD cloud diagram datasets during the period 2011-2017 [9]) the task of manual examining and selecting datasets that can be used for entity linking will become even more challenging. In this work, we argue that data providers will benefit from automating the process of examining datasets and their contents for interlinking.

Recently, methodologies that automatically recommend datasets for entity interlinking have been proposed. They adopt techniques which mainly exploit dataset's instance/schema keywords [10], graph structure [12] or existing links [11] in order to determine the relevancy of datasets for interlinking. Even though a significant number of Web of Data entities are geo-located, to the best of our knowledge, no approach so far makes use of the spatial information available in datasets to recommend relevant datasets for interlinking. According to [8], W3C BasicGeo vocabulary,[2] one of the most well-known spatial vocabularies, is used for assigning coordinates to entities in more than 25% of datasets. In this work, we introduce the exploitation of the spatial information available in datasets for recommending relevant classes for entity interlinking and we examine how geographic approaches can contribute to the problem. We are based on the hypothesis that entity sets (classes) which contain entities that present similar spatial distribution is likely to contain related entities. We argue that comparing the spatial distribution of classes can be identify relevant classes effective-

[1] http://datahub.io/

[2] https://www.w3.org/2003/01/geo/

ly and additionally can reveal relations that cannot be captured by other approaches. For example, it can reveal the topological relation of two, at first sight irrelevant, "Airports" and "Weather Stations" classes. Since many weather stations are located inside airport premises, a data provider might find useful to connect weather stations and their associating airports by a "LocatedIn" relation. Other cases where a geographical approach might be proved useful includes the comparison of datasets that use different languages, schemas or labels to describe related entities.

The goal of our research is to facilitate data providers in the process of discovering Web of Data spatial datasets and classes which may contain related entities with their spatial data. Data providers can then enrich their data by establishing entity links with the identified datasets. Driven by that motivation, we study and propose spatial dataset and class recommendation for entity interlinking methods. Proposed methods will be integrated in a tool that, given as input a spatial entity set, will automatically return a list of Web of Data datasets and their classes that may contain related entities. To fulfill this goal, we firstly parse available in the Web of Data datasets and extract their spatial classes i.e. classes that contain geo-located entities, represented as points. Since the goal of our methodology is to operate at web scale, rather than capturing the actual locations (points) of entities, we propose summarization techniques that capture the spatial distribution of each class. Finally, we apply methods that compare the spatial distributions of the classes in order to identify relevant spatial classes for entity interlinking. In this paper, we also describe our initial experiments that show that our approach can be effective in recommending relevant spatial classes for entity interlinking.

2 State of the Art

Our research addresses the dataset recommendation for entity interlinking problem, which aims at the discovery of Web of Data datasets (or subparts of them) that may contain related entities so as to be used by link discovery methodologies. Typically, the input in dataset recommendation methodologies is a source dataset that is compared against a set of target datasets. The outcome is a (usually ranked) list of relevant (with the source dataset) datasets from the set of target datasets. We identify three main approaches in the existing literature, based on the source of evidence that is used for determining dataset relevancy: a) keyword based b) graph based and c) linkage based approaches. Keyword based approaches measure the string similarity of instance/schema information between datasets. [13] identifies an initial set of candidate datasets by issuing, relevant to the input dataset, keyword queries to a semantic web index (Sig.ma). Then, they rank the initial set of candidate datasets by applying ontology matching techniques that assess the semantic similarity between classes (e.g. string similarity of labels, semantic relations defined in WordNet etc.). Similarly to our approach, they recommend relevant classes for entity interlinking. [14] adopts dataset profiling techniques for characterizing datasets through a set of class labels and they use these profiles to identify schema overlap between datasets. Initially, they identify a cluster of datasets that share schema classes with a given dataset by the help

of a semantico-frequential similarity measure. Then, for each dataset in the identified cluster they compute a dataset relevancy ranking score based on tf*idf cosine similarity. As an additional contribution, their method also returns the mappings between the schema classes across datasets. A dataset recommendation tool, called DRX, which is also based on dataset profiles, was proposed in [15]. Other keyword based methodologies were proposed in [16] and [17]. Graph based approaches compare the similarity of datasets ontology graphs to determine whether two datasets contain related entities. For example, [12] combine Frequent Subgraph Mining techniques to find similarities among datasets. Their approach built on the assumption that "similar datasets should have a similar structure and include semantically similar resources and relationships". They extract frequent subgraphs from RDF datasets and then evaluate the cost of transforming one graph to another. The lower the cost the higher the probability that the two datasets are relevant. Linkage based approaches recommend relevant datasets by using as source of evidence existing links between datasets. [11] develops a Bayesian classifier for ranking datasets according to the probability to define links between URIs of two datasets. The technique uses as evidence of relevance metadata about existing links between all catalogued datasets. [18] uses ranking techniques from social networks for link prediction; the estimation of the likelihood of the existence of an edge between two nodes is based on the already existing links and on the attributes of the nodes. A similar methodology, based on link prediction techniques, was proposed in [19]. We should note that often methodologies use a combination of the above described approaches. For example in [13], additionally to string similarity metrics, they also exploit existing sameAs relations between datasets to determine their relevancy. As pointed earlier, our work is the first that deals with the dataset recommendation for interlinking problem by using as evidence of relevance datasets spatial information.

Another research domain that is closely related to our work is that of dataset summarization. Since capturing analytical information for all Web of Data entities is impractical, Dataset Recommenders usually calculate dataset relevancy based on summarized descriptions of datasets. Dataset profiling is the task of generating a summarized description of a dataset using a set of dataset characteristics [20]. They sketch a taxonomy that discriminates dataset profiles approaches depending on the dataset characteristic they describe. They point that approaches that describe dataset's Domain/Topic [15, 21, 22], Contextual Connectivity [23, 24] or Index/Representative elements [14, 25, 26] can be used for the dataset recommendation for entity interlinking problem. Nevertheless, most works on dataset profiling focus on the profile generation task and do not provide methods for comparing the similarity between dataset profiles. In the geospatial domain, [27] present and compare summarization techniques, distinguished as geometric, space partitioning and hybrid approaches for describing the geographical footprints of point datasets. These summaries are used for answering range and kNN queries. [28] proposes 27 spatial statistics metrics to describe the spatial distribution of feature types and evaluate their discriminative power for the identification of similar feature types. These statistics calculate spatial point patterns (e.g. local intensity, Ripley's K), spatial autocorrelation (e.g. Moran's I) and spatial interactions with other geographic feature types (e.g. count of distinct nearest

feature types). However, as they state, these statistics are mostly descriptive and cannot be used in isolation for effective feature type similarity.

A third related research domain is that of point set similarity which refers to the calculation of a similarity score between two sets of points. In [29] some well know point set distance measures such as Mean, Max, Average, Link and Hausdorff distance are compared regarding their effectiveness for link discovery. These measures calculate the distance between two point sets (in their work a point set represent the vector geometry of an entity) based on the actual point locations and not on summarized descriptions of point sets. [30] applies modified Hausdorff distance measures on Minimum Bound Rectangle (MBR)-based point set summarizations to efficiently calculate similarity on large collections of point sets. Other approaches apply point set similarity techniques to identify similar social network users based on the locations of their activities. [31] proposes and evaluate two distance measures for finding the k-most similar users of a given one: the mutually nearest distance and a QuadTree-based. [32] introduces the Spatio-Textual Point-Set Similarity Join (STPSJoin) query: Given sets of Spatio-Textual objects, each one belonging to a specific type, this query seeks pairs of types that have similar Spatio-Textual objects. Their similarity algorithm uses a similar to Jaccard coefficient metric to measure the overlap of grid based indexed point sets. In this PhD, we examine the applicability of point set similarity methods to the dataset recommendation for entity interlinking problem.

3 Problem Statements and Contributions

The problem of dataset recommendation for entity interlinking can be formulated as follows: Given a source dataset (S) and a set of target datasets (T), identify those Ti ∈ T which may contain related entities for interlinking with S. We intent to contribute to that problem by exploiting the spatial information available in datasets. We extract the spatial classes from each dataset i.e. classes that contain instances for which their geographic location is available, and we compare them to identify the relevant ones for interlinking. We note that we focus on classes which contain instances whose locations are represented as points, excluding thus more complex geometry representations such as lines and polygons. Then, we reformulate the problem: Given a source spatial class (S) and a set of target spatial classes (T), identify those Ti ∈ T which may contain related entities for interlinking with S.

Since this work is the first that exploit the geospatial characteristics of datasets in order to recommend the relevant classes for interlinking, the central research query of this PhD is "How the spatial information of classes can be used for the effective identification of classes that may contain related entities for interlinking". We capture and compare the spatial distribution of classes i.e. the set of the entities' locations which are contained in a class. Our main hypothesis is that "Classes that present similar spatial distribution contain related entities for interlinking". In order to answer our main research question and validate our hypothesis we have to answer the following two questions:

Q1: How to effectively and efficiently summarize the spatial distribution of a spatial class. The goal of our work is to recommend relevant spatial classes for entity interlinking at web scale. A naïve approach would be to capture and operate on the actual entity locations. However, at this scale this seems inefficient and impractical. We, therefore, need to operate on more abstract spatial classes characteristics, like descriptions of the spatial distribution of their entities. We study and evaluate spatial summarization techniques, such as MBRs, spatial indexes and histograms, for their applicability in the dataset recommendation for entity interlinking problem. Proposed spatial summaries should be: a) effective; the description of a class spatial distribution is accurate and b) efficient; summary creation, storage and maintenance costs are low.

Q2: How to compare spatial summaries to effectively determine class relevancy for interlinking. We need metrics that will be applied on the spatial summaries to identify classes that contain related entities. To answer this question we study and evaluate set similarity, distance and probability theory metrics. The proposed metrics should effectively a) identify the relevant pairs of classes and b) rule out the irrelevant classes.

In this research, we argue that our geographical approach may reveal relations between classes that other dataset recommendation for entity interlinking approaches could not identify. For instance, it may identify classes that contain non sameAs but topologically related entities (e.g. Libraries and Universities) or classes that contain sameAs entities described in different languages. Therefore, an additional research question that we target is "Whether a geographical approach can contribute to the dataset recommendation for entity interlinking problem by capturing kinds of relations between datasets that other approaches could not identify". An affirmative answer would be an indication that geographic approaches can be used in combination with other approaches for increasing dataset recommendation for entity interlinking methodologies effectiveness.

The main contributions of this PhD to the research community are:

- We introduce the exploitation of spatial information for recommending Web of Data datasets and classes for entity interlinking and we examine how a geographic approach contributes to the problem
- We propose spatial summarizations techniques and metrics for identifying datasets and classes that may contain related entities
- We provide an easy to use online tool to data providers for the automated and quick discovery of spatial datasets and classes that may contain related entities, facilitating them in the Link Discovery process

4 Research Methodology and Approach

Our overall methodology is mainly divided in five parts: a) spatial class collection b) spatial summaries creation and c) metrics development d) matching algorithm and e) online tool implementation, which are described below.

4.1 Spatial Class Collection

In the first part we identify and collect available Web of Data spatial classes. Collected spatial classes form our basis (database) for the rest parts of our research. Spatial classes are identified and collected automatically according to the following steps:

1. List Web of Data datasets that are provided via a SPARQL Endpoint or as an RDF dump. We acquire this information by automatically parsing CKAN based data catalogs, such as datahub.io. After the execution of this step, we have collected some basic metadata about Web of Data datasets like their name and online resource (e.g. SPARQL Endpoint URL).
2. Identify spatial datasets. Spatial datasets contain spatial entities i.e. entities for which their geographic location, in the form of coordinates, is available. The geographic location of entities is typically described with the use of spatial ontologies. Some well-known spatial ontologies, listed in LOV[3] and LOV4IoT,[4] are W3C Basic Geo, NeoGeo,[5] GeoSPARQL,[6] OrdnanceSurvey,[7] GeoNames[8] and GeoRSS.[9] To identify spatial datasets we issue queries directly to datasets' online resources to check whether they use one of the well-known spatial ontologies. For example, the SPARQL query "ASK {?s <http://www.georss.org/georss/point> ?o}" asks an endpoint whether it uses the point predicate of the GeoRSS ontology. We ask datasets in similar fashion for the remaining spatial ontologies listed above. In this step, we preserve in our database only the detected spatial datasets along with the spatial ontologies that they use. We remind that, as we stated in section 3, we collect only datasets that contain point spatial entities and use the WGS84 coordinate reference system.
3. Identify datasets spatial classes. A dataset may contain one or more classes. Each class contains entities (defined at instance level by the predicate rdf:type) that may be or not spatial. A spatial class is a class that contains spatial entities. Since non spatial classes are irrelevant for our methodology, we maintain only the spatial classes from each dataset. We issue queries directly to datasets online resources to get the list of classes that contain spatial entities using the ontologies that were identified in step 2. For example, the query "SELECT DISTINCT ?class WHERE {?s <http://www.georss.org/georss/point> ?o. ?s <rdf:type> ?class}" returns the list of the dataset's classes that contain entities that use the GeoRSS ontology. To rule out classes that contain very few spatial entities we maintain only those classes that contain 5 or more spatial entities.

[3] http://lov.okfn.org
[4] http://lov4iot.appspot.com/?p=ontologies
[5] http://geovocab.org/
[6] http://www.geosparql.org/
[7] http://data.ordnancesurvey.co.uk/ontology
[8] http://www.geonames.org/ontology/
[9] http://www.georss.org/rdf_rss1.html

The output of the spatial class collection is a list of all the identified Web of Data spatial classes. Each spatial class is described by its URI, the dataset it belongs and the spatial ontology it uses.[10]

4.2 Spatial Summaries Creation

In the second part of our research, we create summaries for the collected spatial classes. We study state of the art spatial summarization techniques and evaluate them for their applicability to the dataset recommendation for entity interlinking problem. Geometric approaches, such as Minimum Bounding Rectangle (MBR), summarize point sets by generating one or multiple geometric shapes that enclose all dataset's points [27]. These techniques are relatively cheap to compute and require low storage, however they do not provide rich dataset descriptions. Space Partitioning approaches, such as spatial indexing, segments the data space into cells. A dataset is summarized by the list of the index cells IDs that are occupied by dataset's points [27]. Compared to the Geometric approaches, they are usually more expensive and require more storage space but they provide richer dataset descriptions. As stated in section 3, Q1 is one of our main research questions. In this PhD, we will develop and compare different spatial summarization techniques.

Currently, we are working on an approach that use both MBR and spatial indexes. For each spatial class we compute and maintain its MBR. Also, we summarize the spatial distribution of the classes trying two spatial indexes: a) a Regular Grid, which partition the global space in equally sized cells (10x10 Km) and b) a QuadTree, which splits space into 4 sub cells recursively according to a criterion (in our case the split criterion is a fixed number of Web of Data spatial entities that occupy a cell, such that high density areas, e.g. city centers, correspond to small sized cells and low density areas, e.g. oceans, correspond to large sized cells). For each indexing method, we use the same index to summarize all spatial classes. For each spatial class, we generate a list of the index cells IDs that intersect with the locations (points) of its entities (we retrieve the entity locations of a class by issuing a SPARQL query to its corresponding endpoint). For example, using the regular grid index (Fig. 1a), the summaries of the triangle and square classes are {2,8,9,11,14,15,16,20} and {6,7,9,13,15,17,20,22} respectively. Using the QuadTree index (Fig. 1b), the summaries of the triangle and square classes are {1,6,11,14,16,17} and {1,5,6,9,13,16,17,18} respectively.

[10] The SPARQL queries that were used for the spatial classes collection and the list of the collected spatial datasets and classes are available in: https://github.com/vkopsachilis/WoDSpatialClassRecommender

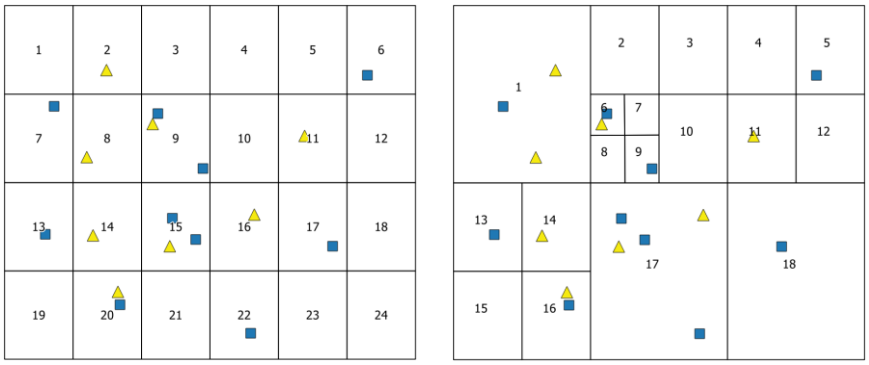

Fig. 1. a) Regular Grid summarization b) QuadTree summarization

The result of this methodology part is that each spatial class is described by its summaries which in our case are its MBR, its Regular Grid-based and its QuadTree-based summary.

4.3 Metrics Development

This part refer to the development of metrics that will be applied on the class spatial summaries to compute their similarity. If two classes present similar spatial summaries (which as stated earlier represent classes spatial distributions), then it is likely that they contain related entities for interlinking. Proposed metrics should have the discriminative power to accurately and precisely identify relevant classes. We study and evaluate state of the art metrics from different domains (e.g. set similarity, probability theory) for their suitability to the dataset recommendation problem. For our first experiments (section 5) we adapted and evaluated four metrics. We use the following notation: e represents the total number of the index cells (e.g. QuadTree) contained in a given geographic area; a and b the number of the class A and B summary cells respectively contained in a given geographic area; and c the number of the cells that the two classes have in common. In the example of Fig 1b, these number would be e=18, a=6 (triangles), b=8 (squares) and c=4.

- *Jaccard Similarity:* calculates the number of the common cells divided by the size of the union cells of two classes. *JS* returns values between 0 and 1. Values approaching 1 means high similarity while values approaching 0 means low similarity.

$$JS = \frac{c}{a + b - c}$$

- *Overlap Coefficient*: calculates the number of the common cells divided by the cells number of the smallest class summary. *OC* returns values between 0 and 1.

Values approaching 1 means high similarity and values approaching 0 means low similarity.

$$OC = \frac{c}{\min(a, b)}$$

- *HyperGeometic Distribution Cumulative Probability*: estimates the probability of existing *c* or more common cell when class A occupies *a* cells and class B occupies *b* cells in a given area covered by *e* cells. *HG* returns values between 0 and 1. Low probability values imply that is unlikely for two (random) classes to have *c* or more common cells, therefore they must be related.

$$P(X \geq c) = \sum_{c}^{min(a,b)} P(X = c) \text{ where } P(X = c) = \frac{\binom{a}{c}\binom{e-a}{b-c}}{\binom{e}{b}}$$

- *Independent Events Probability Ratio:* calculates the ratio of the common cells to the number of the expected common cells (c_{exp}). IR values have no upper limit. High IR values imply that two classes are likely to contain related entities. Expected common cells is the number of the common cells that two classes would have if they were not related (independent). We calculate c_{exp} by adapting the independent event probability formula: the probability for two not related (A and B) classes to have common cells in a given area (c_{exp}/e) is the product of the probabilities of class A (a/e) and class B (b/e) (i.e. the number of cells that a class occupies in a given area).

$$IR = \frac{c}{Cexp} \text{ where } Cexp = \frac{ab}{e}$$

4.4 Matching Algorithm

The goal of the matching algorithm is to identify spatial classes that may contain related entities and thus are relevant for interlinking. Below, we sketch the execution order of the algorithm:

Input: A source spatial class and a set of target spatial classes (formed by the set of already collected classes in the spatial class collection part). The source class can be selected from the list of the already collected classes or might be a new one. In the latter case, we create the summaries for the new class according to the methodology described in the spatial class summarization part.

1. Filter out as irrelevant, target classes with non-overlapping MBR with the source class in order to identify an initial set of candidate classes.
2. Compare the spatial index summary (e.g. QuadTree-based) of the source class with the respective summary (i.e. QuadTree-based) of each candidate class in the overlapping MBR area by calculating the values of the metrics described in section 4.3

3. Determine whether the source class is relevant for interlinking with a candidate class by checking if the calculated metrics values for this pair of classes satisfy some criteria e.g. exceed a metric threshold.

Output: A list of the relevant classes for interlinking with the source class. Part of future work is to provide a ranked list of relevant classes.

4.5 Online Tool

The last part of our work refer to the development of an online tool that will be the entry point to the matching algorithm, allowing data providers to easily discovery relevant spatial classes for entity interlinking.

5 Preliminary Results

At this point of our research we have identified and created summaries for about 20700 spatial classes from 57 different datasets provided via SPARQL endpoints. The three datasets that contain the most spatial classes (collected as described in 4.1) are DBpedia, an online repository of links between Knowledge Bases called LinkLion,[11] and a service that delivers RDF based descriptions of Web addressable resources called URIBurner[12] (15488, 898 and 544 spatial classes respectively).

We conducted a first experiment to assess the effectiveness of the developed spatial summaries and metrics for discovering classes that contain related entities on a randomly selected set of 100 spatial classes. We examined them manually to identify pairs of classes that contain related entities. We found that 20 pairs of classes contain related entities while the rest pairs (4930) are irrelevant for interlinking. We, then applied our matching algorithm using as input each time a different source class and comparing it with the rest sample classes, thus resulting in 100 runs and 4950 pair comparisons in total. For each pair of classes comparison, we calculated all possible metric/summary values (e.g. Jaccard Similarity for Regular Grid summaries, Jaccard Similarity for QuadTree summaries, Overlap Coefficient for Regular Grid summaries and so on) and we manually defined some metrics thresholds to determine whether a pair of class should be returned as relevant for interlinking or not. To assess the effectiveness of the various metric/summary and thresholds combinations we calculated their Recall and Precision for all runs. Recall calculates the number of the correctly identified relevant pairs of classes divided by the number of the total relevant pairs of classes in our sample (that is 20), while precision calculates the number of the correctly identified pairs of classes divided by the number of total identified pairs of classes.[13]

[11] http://www.linklion.org/

[12] http://uriburner.com/

[13] Sample classes, ground truth pairs of relevant classes and analytical results for the experiments are available in https://github.com/vkopsachilis/WoDSpatialClassRecommender/tree/master/Experiments

The results of the first experiment showed that the 10X10Km Regular Grid is ineffective (averaging 0.03 precision and 0.71 recall for all the regular gird/metric combinations that we tested) for identifying relevant datasets for interlinking. Of course, using a Regular Grid with smaller sized cells would increase its precision but this would explode storage requirements and computational costs. On the other hand, QuadTree summaries proved effective for identifying relevant classes, since they averaged 0.56 precision and 0.74 recall for all QuadTree/metric combinations that we tested. Concerning the metrics, the HyperGeometic Distribution Cumulative Probability (HG) and Independent Events Probability Ratio (IR) achieved better F-scores than the Jaccard Similarity (JS) and Overlap Coefficient (OC) metrics for the various thresholds that we tested. The best score for each metric when applied on QuadTree summaries is: HG scored 0.80 recall and 0.51 precision when the threshold set to HG<0.00001 and IR scored 0.95 recall and 0.54 precision when the threshold set to IR>5. JS scored 0.20 recall and 1 precision when the threshold set to JS>0.2 and OC scored 0.80 recall and 0.23 precision when the threshold set to OC>0.1. The above results indicate that applying HG and IR metrics on QuadTree summaries is the most effective combination that we tested so far for recommending relevant classes for entity interlinking.

We conducted a second experiment using a different sample to confirm the methodology effectiveness. To form our sample we selected 25 classes and we took care to include pairs of relevant classes that belong to different datasets and pairs of relevant classes with different class labels (e.g. "Cathedrals" and "PlaceOfWorship"). In this sample 28 pairs of classes were identified as relevant and the rest 272 pairs as irrelevant. Similar to the first experiment, we applied our matching algorithm using as input a different source class each time and we compare it with the rest sample classes, thus resulting in 25 runs and 300 pairs comparison in total. For each pair of class comparison, we calculated metric values only for QuadTree summaries (since regular grid was proved ineffective) and we used the metrics thresholds from the first experiment. Similarly to the first experiment, we were based on the recall and precision metrics to evaluate the methodology effectiveness. The results of the second experiment confirmed the finding that HG and IR metrics perform better than the JS and OC metrics. HG scored 0.79 recall and 0.58 precision when the threshold set to HG<0.00001 and IR scored 0.64 recall and 0.95 precision when the threshold set to IR>5. JS scored 0.07 recall and 1 precision when the threshold set to JS>0.2 and OC scored 0.71 recall and 0.65 precision when the threshold set to OC>0.1. The results of our initial experiments indicate that the exploitation of the dataset's spatial information can contribute to the dataset recommendation for entity interlinking problem. Part of this research we will dedicated in the research and development of more effective and efficient summaries and metrics.

6 Evaluation Plan

In the future, we plan to perform more experiments that would enables us to draw safer conclusion regarding the effectiveness of our approach for recommending rele-

vant class for interlinking. For instance, in subsequent experiments we will formulate more unbiased sample class sets by including classes from more datasets and taking into account factors like class size and geographical extent diversity. A method that will return a ranked list of relevant datasets will be integrated and evaluated using appropriate metrics such as Precision@N. Moreover, Data Mining techniques can be used for determining metrics thresholds. As our research evolves, we will be based on a bigger training set of classes that will help in the methodology optimization. Finally, we will compare our approach with other state of the art works, particularly for answering the question of how a spatial approach differentiates, regarding the kind of relations that can identify, from other dataset recommendation for entity interlinking approaches.

7 Conclusion

Our research focuses on recommending Web of Data spatial datasets and classes that can be used for entity interlinking. To achieve this, we propose a methodology that exploits the spatial information available in datasets. We are based on the assumption that classes that present similar spatial distribution is likely to contain related entities. We identify Web of Data spatial classes, we capture their spatial distributions and we compare them to identify the relevant classes for interlinking. Our initial experiments indicate that our spatial approach can contribute to the problem. We continue our research for the development of more effective and efficient summaries and metrics and for drawing conclusions about what kind of insights to the dataset recommendation problem can a spatial approach provide.

Acknowledgements. This research is being completed under the supervision of Ass. Prof. Michail Vaitis and is being supported by the funding program "YPATIA" of University of Aegean.

References

1. Berners-Lee, T., Linked Data, https://www.w3.org/DesignIssues/LinkedData.html, last accessed 2018/03/05.
2. Heath, T., Bizer, C.: Linked data: evolving the Web into a global data space. In: Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers (2011).
3. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the Web of data. In: Proceedings of the 8th International Semantic Web Conference, pp. 650–665. Springer, Heidelberg (2009).
4. Ngomo, A.C.N., Auer, S.: LIMES - A time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, pp. 2312–2317. AAAI Press (2011).
5. Nentwig, M., Hartung, M., Ngomo, A.C.N., Rahm, E.: A survey of current Link Discovery frameworks. In: *Semantic Web*, vol. *8*(3), pp. 419–436. (2017).
6. The Linking Open Data cloud diagram, http://lod-cloud.net/,last accessed 2018/03/05.

7. LODStats, http://stats.lod2.eu/, last accessed 2018/03/05.
8. State of the LOD Cloud 2014, http://lod-cloud.net/state/state_2014/, last accessed 2018/03/05.
9. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Mika P. et al. (eds) The Semantic Web – ISWC 2014. Lecture Notes in Computer Science, vol. 8796, pp. 245–260. Springer, Cham (2014).
10. Nikolov, A., D'Aquin, M.: Identifying relevant sources for data linking using a semantic web index. In: *WWW2011 Workshop: Linked Data on the Web,* (2011).
11. Leme L.A.P.P., Lopes G.R., Nunes B.P., Casanova M.A., Dietze S.: Identifying Candidate Datasets for data interlinking. In: Daniel F., Dolog P., Li Q. (eds) Web Engineering. ICWE 2013. Lecture Notes in Computer Science, vol 7977. Springer, Berlin, Heidelberg (2013).
12. Emaldi, M., Corcho, O., López-de-Ipiña, D.: Detection of related semantic datasets based on frequent subgraph mining. In: IESD@ISWC. (2015).
13. Nikolov, A., D'Aquin, M., Motta, E.: What should i link to? Identifying relevant sources and classes for data linking. In: Proceedings of the 2011 joint international conference on The Semantic Web, pp. 284–299. Springer Berlin, Heidelberg (2012).
14. Ben Ellefi, M. Ben, Bellahsene, Z., Dietze, S., Todorov, K.: Dataset recommendation for data linking: An intensional approach. In: Sack H., Blomqvist E., d'Aquin M., Ghidini C., Ponzetto S., Lange C. (eds) The Semantic Web. Latest Advances and New Domains. ESWC 2016. Lecture Notes in Computer Science, vol. 9678. Springer, Cham (2016).
15. Arturo, A., Caraballo, M., Nunes, B. P., & Casanova, M. A.: DRX: A LOD dataset interlinking recommendation tool. (2015).
16. Mehdi, M., Iqbal, A., Hogan, A., Hasnain, A., Khan, Y., Decker, S., Sahay, R.: Discovering Domain-Specific Public SPARQL Endpoints: A Life-Sciences Use-Case. In: Proceedings of the 18th International Database Engineering & Applications Symposium, pp. 39–45. ACM, New York (2014).
17. Martins, Y.C., da Mota,F., Cavalcanti M.C: DSCrank: A Method for Selection and Ranking of Datasets. Garoufallou E., Subirats Coll I., Stellato A., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham. (2016).
18. Lopes G.R., Leme L.A.P.P., Nunes B.P., Casanova M.A., Dietze S. Recommending Tripleset Interlinking through a Social Network Approach. In: Lin X., Manolopoulos Y., Srivastava D., Huang G. (eds) Web Information Systems Engineering – WISE 2013. Lecture Notes in Computer Science, vol 8180. Springer, Berlin, Heidelberg (2013).
19. Liu, H., Wang, T., Tang, J., Ning, H., Wei D.: Link prediction of datasets sameAS interlinking network on web of data. In: 3rd International Conference on Information Management, (2017).
20. Ben Ellefi, M., Bellahsene, Z., Breslin, J. G., Demidova, E., Dietze, S., Szymaski, J., Todorov, K. (2016). RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. In: Semantic Web J. (2017).
21. Lalithsena, S., Hitzler, P., Sheth, A., Jain, P.: Automatic Domain Identification for Linked Open Data. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). (2013).
22. Fetahu, B., Dietze, S., Nunes, B. P., Taibi, D., Casanova, M. A. Generating structured profiles of linked data graphs. In: Proceedings of the 12th International Semantic Web Conference (ISWC). Springer (2013).
23. Mountantonakis, M., Allocca, C., Fafalios, P., Minadakis, N., Marketakis, Y., Lantzaki, C., Tzitzikas, Y.: Extending VoID for expressing connectivity metrics of a semantic ware-

house. In: 1st International Workshop on Dataset Profiling & Federated Search for Linked Data. (2014).
24. Wagner, A., Haase, P., Rettinger, A., Lamm, H.: Entity-based data source contextualization for searching the web of data. In: The Semantic Web: ESWC 2014 Satellite Events. (2014).
25. Böhm, C., Lorey, J., Naumann, F.: Creating voiD descriptions for Web-scale data. In: Journal of Web Semantics, vol. 9(3), pp. 339–345. (2011).
26. Hasnain, A., Zainab, S., Hasnain, A., Hogan, A.: SPORTAL: Profiling the Content of Public SPARQL Endpoints. In: International Journal on Semantic Web and Information Systems, vol. 12(3), pp. 134–163. (2016).
27. Kufer, S., Henrich, A.: Hybrid Quantized Resource Descriptions for Geospatial Source Selection. In: International Conference on Information and Knowledge Management, pp. 17–24. (2014).
28. Zhu, R., Hu, Y., Janowicz, K., McKenzie, G.: Spatial signatures for geographic feature types: examining gazetteer ontologies using spatial statistics. In: Transactions in GIS, *vol. 20*(3), pp. 333–355. (2016).
29. Sherif, M. A., Ngomo, A. N.: A Systematic Survey of Point Set Distance Measures for Link Discovery. In: *Semantic Web*, vol. *1*(2013), pp. 1–5. (2014).
30. Adelfio, M. D., Nutanong, S., Samet, H.: Similarity search on a large collection of point sets. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '11, pp. 132-141. (2011).
31. Kanza, Y., Kravi, E., Safra, E., Sagiv, Y.: Location-Based Distance Measures for Geosocial Similarity. In: ACM Transactions on the Web, vol. 11(3). ACM, NewYork (2014).
32. Efstathiades, C., Belesiotis, A., Skoutas, D., Pfoser, D.: Similarity Search on Spatio-Textual Point Sets. In: 19th International Conference on Extending Database Technology. (2016).