Focused Exploration of Geospatial Context on Linked Open Data

Thomas Gottron[1], Johannes Schmitz[1], Stuart E. Middleton[2]

[1] Institute for Web Science and Technologies, University of Koblenz-Landau, Germany
{gottron,schmitzj}@uni-koblenz.de
[2] University of Southampton IT Innovation Centre, UK
sem@it-innovation.soton.ac.uk

Abstract The Linked Open Data cloud provides a wide range of different types of information which are interlinked and connected. When a user or application is interested in specific types of information under time constraints it is best to explore this vast knowledge network in a focused and directed way. In this paper we address the novel task of focused exploration of Linked Open Data for geospatial resources, helping journalists in real-time during breaking news stories to find contextual geospatial information related to geoparsed content. After formalising the task of focused exploration, we present and evaluate five approaches based on three different paradigms. Our results on a dataset with 425,338 entities show that focused exploration on the Linked Data cloud is feasible and can be implemented at very high levels of accuracy of more than 98%.

1 Introduction

The Linked Open Data (LOD) cloud is a rich resource for geospatial information. This is reflected in the LOD cloud diagram with geospatial Linked Data resources representing a large part of the overall cloud (cf. Fig. 1). Thanks to the network structure of Linked Data it is possible to explore both the semantic and geographic context, starting from a known location entity and following references through to other entities.

Applications can utilise this rich resource when providing geospatial context information to end users. For example in the REVEAL project[3] we provide real-time news room social media analytics to journalists, making use of Linked Data resources to both augment real-time situation assessments, such as a breaking news story of a major flooding event, and provide context to assist the verification and analysis of social media content behind these news stories such as official flood risk assessment data.

In practice, however, exploring the semantic neighbourhood of a location entity involves following multiple links and dereferencing the URIs representing the corresponding entities. As the outdegree of Linked Data nodes can be high—during our investigations we typically observed between 15 and 100 outgoing links—this involves a potentially large network communication and data transmission overhead in collecting the data. The resulting latency in collecting this information might be too much for time critical applications or use cases where network bandwidth is limited (e.g. when

[3] http://revealproject.eu

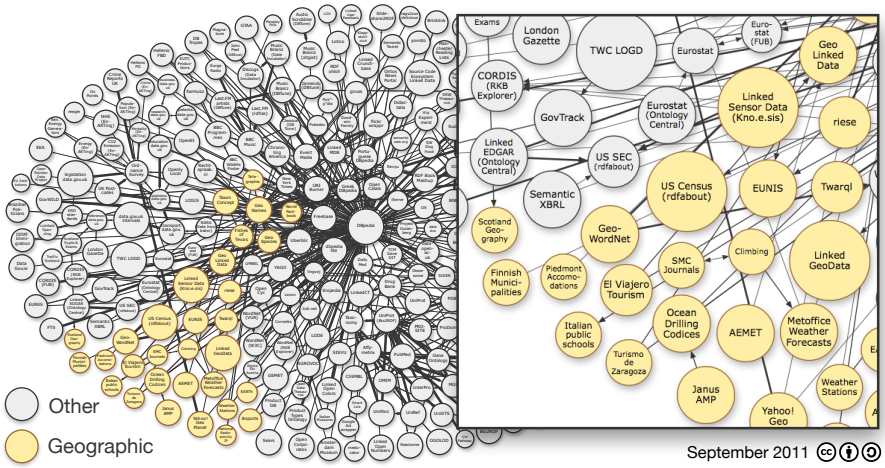

Figure 1. The Linked Open Data cloud diagram with DBpedia in the centre [6]. The data sources in the geographic domain are zoomed and highlighted in yellow.

systems aim for near real time analytics of data streams or when end users explore geospatial data on mobile devices). This becomes even more an issue when only a few of the links lead to relevant contextual information such as geospatial resources. Typical solutions to this challenge involve operating on local and aggregated copies of relevant Linked Data which are maintained in data caches [11]. Caching approaches are not optimal however for applications where resources are updated frequently or where relevant entities are not known in advance [5].

In this paper we address the question of whether it is possible to perform a focused and directed exploration of LOD for geospatial context. To the authors' knowledge this challenge has not been investigated before. Following our geospatial use case we attempt to explore only those links starting from a seed entity which lead to geospatial resources. This means we prioritise the outgoing links of a location entity to decide which URIs to dereference first. We base this prioritisation on information encoded in the semantic links, i.e. the predicates which lead to the referenced LOD entities. Our hypothesis is that we can learn from the types of references to other entities which entities are of a geospatial type themselves and are likely to have information about a geo-coordinate.

We investigate five approaches based on three different paradigms: (a) one approach using the semantics of RDFS schema definitions for predicates, (b) two variations of supervised classifiers which use predicate types as features and (c) two approaches inspired by Information Retrieval (IR) techniques on the descriptiveness of terms. We have implemented all five approaches and run an evaluation using a real world dataset with 425,338 entities to benchmark their predictive performance.

This paper proceeds with a detailed description and formalisation of the task of focused exploration in Sections 2 and 3. We then describe our five prediction models, explain our design decisions and implementation choices. Afterwards we evaluate the

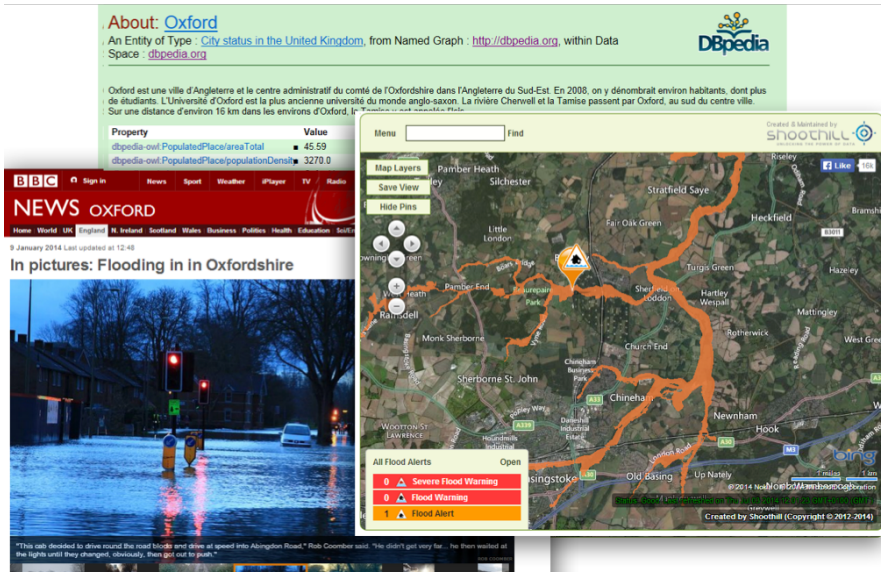

Figure 2. Oxford flooding use case and Linked Data resources. Images courtesy of BBC news, DBpedia and the UK Environment Agency.

approaches and discuss their relative performance in Section 5. In Section 6 we look at related work, before concluding with a summary and an outlook for future work.

2 Exploring Geospatial Context on the LOD Cloud

A significant share of the LOD cloud deals with geospatial information (cf. Fig. 1). The entities in this part of the cloud typically provide a geolocation in the form of coordinates and information related to the entity itself. This information is valuable for many use cases and applications.

For example during 2014 there was a major flooding event in Oxford, UK. This flooding was well documented by journalists in the UK, e.g. at the BBC. The REVEAL project provides real-time geoparsing of location data [7] from large volumes of social media content (e.g. Twitter streams) visualised using situation assessment maps such as flood incident maps. Locations are extracted from social media reports and annotated with linkedgeodata.org URIs. This provides us with initial Linked Data entities for flooded locations from which we can do a first hop of exploration to DBpedia, and follow-on hops to links containing relevant contextual geospatial information such as regional population data, socio-economic data for impacted regions and UK environment flood risk assessment maps. Figure 2 contains some screenshots from BBC News relating to Oxford flooding, DBPedia Linked Data for Oxford and Open Data resources such as live flood alert maps from the UK environment agency websites.

Our motivation is to explore the context of each location using semantic links. The links to the related items are typed to model the semantic relations and can imply that

certain entities are locations. However, on the LOD cloud one can never be certain that the semantics hold. Furthermore, the semantic declaration of a location entity does not necessarily imply the availability of geo-coordinates. At the same time the outdegree of the nodes makes it difficult to follow all links in time critical applications or in scenarios where bandwidth is a limiting factor. This motivates the question of alternative approaches for prioritising or filtering links to related entities and to perform a focused exploration of Linked Data.

3 Task Definition: Focused Exploration on Linked Data

In the context of our geospatial use case, the task of focused exploration on Linked Data can be be formalised as follows. We have got a set of entities E modelled on the Linked Data cloud and represented by URIs. Information about the entities is expressed as triple statements (s, p, o) where $s \in E$, p is a predicate denoting a specific property or relation of the entity (also expressed as a URI) and o is the object of this relation and can be a literal value or another entity URI. The set of all statements is denoted by R.

Among the entities in E there are some which represent locations and have an associated geo-coordinate. In our setting these are the only location entities which are of relevance, as they are the only ones which can be located on a map. The W3C recommendation to represent locations is by using the WGS84 standard to define a position via latitude and longitude. Hence, we define a subset $L \subset E$ of relevant *location entities* which provide WGS84 coordinates. Formally, we can define L as:

$$L := \{x \in E \mid (x, \mathsf{wgs84{:}lat}, \textit{latitude}) \in R \wedge (x, \mathsf{wgs84{:}long}, \textit{longitude}) \in R\} \quad (1)$$

In our scenario we are now facing a situation in which we are provided with an entity $x \in L$. Furthermore, we have access to all statements $(x, p, o) \in R$, where the entity x appears in the subject position. The task we intend to solve is to predict which of the objects in the set $\{o : (x, p, o) \in R\}$ are also elements of L, i.e. do provide a geo-coordinate. This focused exploration task can be formalised from two different points of view: (a) as a classification task and (b) as a ranking task.

Formalising the task as a *classification task* means that we have to assign a class label l to each object URI $o \in E$. The label will be $l = 1$ for the objects which are predicted to provide a geo-coordinate and 0 for those which are predicted not to provide a coordinate. This is a simple binary decision which needs to be made on the basis of some features and information we have available about the URI for o (e.g. the type of predicate used to link to it). In an application setting, the use of the predicted labels would be to dereference only those objects o which have a label of $l = 1$.

When formalising the task as a *ranking task*, the setting is slightly different. Instead of selecting the presumably relevant objects, we sort them from the object most likely to provide a location to the object least likely. This means we derive a ranking $(o_1, o_2, \ldots, o_n)$ of the objects. This formalisation is favorable for application scenarios where it is possible to dereference and retrieve the information about the objects going through the ranked list in a top down way. The benefit of the ranked list is that this approach can be pursued until some limiting criterion is reached (e.g. time, number of URIs, used bandwidth, etc.).

4 Approaches

To solve the task we described in Section 3, we are considering five approaches based on three different paradigms: (a) making use of the semantics in schema information for RDFS vocabularies, (b) supervised machine learning and (c) Information Retrieval inspired approaches making use of the discriminativeness of RDF predicates.

4.1 Schema Semantics

Many of the vocabularies used to model data on the LOD cloud are based on RDFS (or even more expressive languages). This means we can find schematic information about the predicates and RDF types used to model the entities and their relationship. One such type of information is rdfs:range which provides the type of objects referenced by predicates.

The *schema semantics* approach makes use of this information in order to be able to prioritise fetching URIs which have been referenced by a predicate with an rdfs:range of RDF types related to locations. To this end, we collected schema information about predicates and checked their rdfs:range definitions. If the types defined there semantically represent locations (e.g. dbpedia:Place and all it subclasses) we consider them as relevant, otherwise as not relevant.

Formally, this provides us with a set of predicates P_L for which we can infer that the objects they link to are locations. Accordingly, we declare an object o as relevant (i.e. assign a label of $l = 1$) if we observe a statement $(x, p, o) \in R$, where $p \in P_L$. When operating in the ranking setting, we simply sort the objects in a way, that all objects with a label $l = 1$ are ranked higher than the ones with a label $l = 0$.[4]

4.2 Supervised Machine Learning

In this case we tackle the task of focused exploration as a learning problem. The learning task is based on observation of predicates used to express relations to object entities and observations of whether the object entities actually did or did not provide geo-coordinates. Using statistical learning methods, we then infer which combinations of predicates are more likely to lead to location entities in L than others.

As features of the objects we use types of predicates that are used to reference them. These represent binary features of predicates being used or not used to refer from a subject URI to an object URI. To formalise the approach, let us define the set of predicates as $\{p_1, p_2, p_3, \ldots, p_m\}$. Assume now, we have got an object x which is referenced by some predicates p_i with $i \in I \subset \{1, 2, 3, \ldots, m\}$. Accordingly we can represent an object x by a feature vector $(\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3, \ldots, \mathsf{p}_m)$ where $\mathsf{p}_i = 1$ if $i \in I$ and $\mathsf{p}_i = 0$ if $i \notin I$.

We employed a Naive Bayes classifier which can deal with large datasets, high number of features and the binary categorial type of the features. We investigated two variations. When classifying an object, the first variation of the Naive Bayes classifier is using information about both: presence and absence of predicates. This means we

[4] Among the objects with the same label we use a random ordering.

are also able to infer information about the relevance of the object if certain predicates are *not* used to refer to it. The second variation makes use only of the predicates actually observed for a concrete object. Formally, we can distinguish the two different approaches by their underlying probabilities:

$$P_{all}\left(x \in L \,|\, (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right) \propto \prod_{i \in I} P(\mathsf{p}_i = 1 \,|\, x \in L) \cdot \prod_{i \notin I} P(\mathsf{p}_i = 0 \,|\, x \in L) \cdot P(x \in L) \quad (2)$$

$$P_{observed}\left(x \in L \mid (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right) \propto \prod_{i \in I} P(\mathsf{p}_i = 1 \mid x \in L) \cdot P(x \in L) \quad (3)$$

The probabilities are estimated from training data using a Maximum Likelihood estimator and Laplace smoothing [4].

When addressing the classification variation of our task, we can compute for a given object x both probabilities $P\left(x \in L \mid (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right)$ and $P\left(x \notin L \mid (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right)$ and assign x to the category with the higher probability. For the ranking task we need to combine both probabilities into a single ranking score. In this case we use the odds $O\left(x \in L \mid (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right) = \frac{P\left(x \in L | (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right)}{P\left(x \notin L | (\mathsf{p}_1, \ldots, \mathsf{p}_m)\right)}$ as score for ranking.

4.3 Information Retrieval Inspired Approach

The last two approaches are heuristics inspired by *tf-idf* measures from the domain of Information Retrieval. The ingredients here are twofold. First, we want to measure how often a predicate is used to link to a relevant URI. For this we use a frequency that mimics the term frequency (*tf*) measure in *tf-idf*. Second, we want to distinguish how discriminative a predicate is on a dataset level. A predicate which appears very frequently and references to nearly all objects cannot discriminate very well. This follows the idea of the inverse document frequency (*idf*) of terms in Information Retrieval systems.

We define the *predicate relevance frequency (prf)* measure for a predicate as:

$$prf(p) = c(p, L) \quad (4)$$

where $c(p, L)$ gives the number of links with predicate p which lead to a relevant object URI. The higher the *prf* value of a predicate p, the more often p has lead to a relevant object. This should indicate the importance of the predicate p for finding relevant objects. An alternative to *prf* is to normalise the frequency in order to remove a bias towards very frequent predicates. This leads to the *predicate relevance ratio (prr)*:

$$prr(p) = \frac{c(p, L)}{c(p, *)} \quad (5)$$

where the normalisation term $c(p, *)$ gives the overall number of links with predicate p.

As second measure we define the *inverse predicate frequency (ipf)* for p as:

$$ipf(p) = log\left(\frac{c(*,*)}{c(p,*)}\right) \tag{6}$$

where $c(p,*)$ is again the overall number of links with predicate p and $c(*,*)$ gives the entire number of links in the training dataset. The higher the *ipf* value of a predicate p, the less often it is used to reference an object URI. The more often the predicate is used, instead, the lower the *ipf* value. In particular, a predicate that is used for all objects is not discriminative at all and will get an *ipf* value of 0.

We then combine these measures into *prf-ipf* and *prr-ipf* weights for a predicate p:

$$w_{prf\text{-}ipf}(p) = prf(p) \cdot ipf(p) \qquad w_{prr\text{-}ipf}(p) = prr(p) \cdot ipf(p) \tag{7}$$

When ranking an entity it is assigned a score ρ which corresponds to the aggregated *prr-ipf* or *prr-ipf* weights of all its predicate features. For the classification interpretation of the task, we need to define a threshold θ for the score. All entities with a score above θ will be considered as relevant, all entities with a score below θ are labeled as irrelevant. Also the threshold is derived from the training data. To this end, we compute the mean relevance score ρ_{rel} of all actually relevant objects in the training data. In the same way we determine the mean relevance score ρ_{irrel} of irrelevant data. This two values serve as reference points for a simple nearest centroid classifier, which corresponds to using a final threshold $\theta = \frac{\rho_{rel}+\rho_{irrel}}{2}$.

5 Evaluation

In this section we address the question of how well the approaches presented in Section 4 perform in our focused exploration task for geospatial entities on Linked Data.

5.1 Dataset

For evaluation purposes we constructed a dataset of entities with geolocations and the entities they refer to. We started from the owl:sameAs links between LinkedGeoData and DBPedia which provides us with 99,951 entities of locations in DBPedia. These entities constitute our seed dataset. For each of these entities we want to explore the context for finding further entities with geo-coordinates.

In the seed data we observed a total of 1,728,633 outgoing links which refer to URIs. We then excluded all references based on owl:sameAs links, as they would not lead to *new* information. Likewise we excluded schema information defined by rdf:type statements. Furthermore, we filtered out some types of links (foaf:homepage, dbpedia:wikiPageExternalLink, foaf:isPrimaryTopicOf, foaf:depiction) which lead to URIs which do not represent Linked Data entities but lead to HTML Web documents or other file formats on the Web. This left us with 425,338 distinct URIs of entities which were candidates for an exploration[5]. As features of the objects we considered all predicates

[5] Please note that the number of links leading to these entities is even higher, as several entities are referenced by more than one link predicate.

which appeared in incoming links to these 425,338 URIs. We only removed rare predicates, i.e. which appeared with less than 10 of the URIs. This lead to a total of 353 different predicates which served as features.

All of the 425,338 URIs were dereferenced and we retrieved the Linked Data description of the modeled entities. As part of the descriptions we identified a total of 128,171 entities with geo-locations. Thus, for this explored data we were able to provide a gold standard of relevant objects based on these descriptions.

5.2 Evaluation Methodology and Metrics

We address the task of focused exploration in both possible ways of interpretation: as ranking task and as classification task. For the ranking task we derived receiver operating characteristic (ROC) curves and computed the area under curve (AUC) as a metric to evaluate effectiveness. For the classification task we present the results in the form of a confusion matrix and compute Precision, Recall, F1 and Accuracy. As the machine learning and the Information Retrieval models need training data we perform a cross validation. Specifically we use a 10-times 10-fold cross validation approach. The results we present in the following are average values over the ten iterations.

5.3 Results

We start to look at the performance under the ranking aspect of the focused exploration task. Figure 3 shows the ROC curves of all considered approaches. We can see that the trained models perform far better than the model based on schema semantics. The curves of the Naive Bayes (NB) models and the *prf-ipf*-model are all relatively close to each other indicating a similar performance. Furthermore all curves start with a very steep inclination and than flatten out at a value close to one. To better see the minor differences we have enlarged the top left corner of the plot. There we can see that the Naive Bayes models are slightly better than the Information Retrieval inspired models. Among the two it is difficult to declare a better approach as the ROC curves cross and overlap.

When considering focused exploration as a classification task, we observe a similar behaviour. In Table 1 we see confusion matrices for all of the approaches[6]. The approach based on schema semantics performs relatively poor in identifying relevant objects. The total number of true positive classification is far lower than for all other approaches. The approaches making use of a Naive Bayes classifier perform very well. Most instances are classified correctly, the rate of false classifications is below 2% on our dataset. Also *prf-ipf* an *prr-ipf* show a good performance. Regarding the types of mistakes made, the two approaches have a opposing behaviour. While *prf-ipf* is less prone to erroneously label an irrelevant object as relevant (fewer false positives), *prr-ipf* misses less relevant documents (fewer false negatives).

Table 2, finally summarises the performance of the approaches with the aggregated average measures for Recall, Precision, F1, Accuracy and AUC over the full ten iterations of the 10-fold cross validation. For each of the measures we have marked the

[6] The confusion matrices were chosen randomly from one the ten iterations in the cross-validation. However, the numbers are very stable over all iterations.

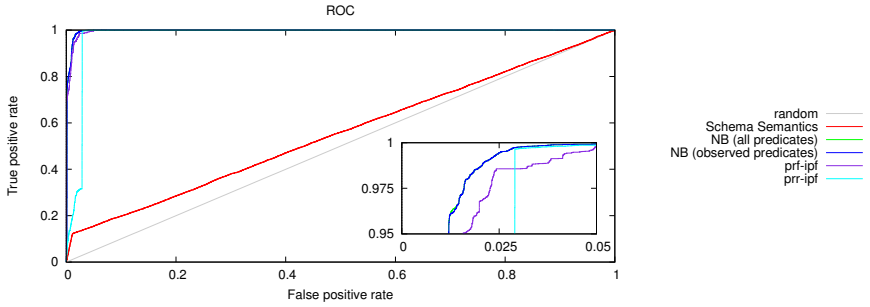

Figure 3. ROC curves of all approaches for the ranking task. The upper left corner of the plot is enlarged to illustrate details.

Table 1. Confusion Matrices of Approaches

Schema Semantics

		Ground truth: Relevant	Ground truth: Irrelevant
Class	Relevant	15,227	3,528
	Irrelevant	112,944	293,639

NB (all predicates)

		Ground truth: Relevant	Ground truth: Irrelevant
Class	Relevant	126,993	6,831
	Irrelevant	1,178	290,336

NB (observed predicates)

		Ground truth: Relevant	Ground truth: Irrelevant
Class	Relevant	127,446	7,624
	Irrelevant	725	289,543

prf-ipf

		Ground truth: Relevant	Ground truth: Irrelevant
Class	Relevant	109,107	2,753
	Irrelevant	19,064	294,414

prr-ipf

		Ground truth: Relevant	Ground truth: Irrelevant
Class	Relevant	127,818	10,454
	Irrelevant	353	286,713

best performance in bold. Furthermore, we marked the results where we had a significant improvement over the second best method at confidence level of $\rho = 0.01$. The aggregated values basically confirm the observations made above. In general, when considering the measures F1, Accuracy and AUC, the Naive Bayes classifier making use of all predicates performs best. In application scenarios, where a high Recall is of importance, instead, the *prr-ipf* approach achieves the best results with more than 99.7%. When focusing on Precision, *prf-ipf* performs best and demonstrated the highest values. More than 97% of the objects predicted to have geo-coordinates actually did provide such information. In a setting where we want to focus on promising items this might be the kind of performance the end user is looking for.

Table 2. Average performance of approaches († indicates significant improvements at confidence level $\rho = 0.01$)

Method	Recall	Precision	F1	Accuracy	AUC
Schema Scemantics	0.1188	0.8119	0.2073	0.7262	0.5552
NB (all predicates)	0.9906	0.9491	† **0.9694**	† **0.9812**	**0.9970**
NB (observed predicates)	0.9943	0.9436	0.9683	0.9804	0.9968
prf-ipf	0.8512	† **0.9754**	0.9091	0.9487	0.9958
prr-ipf	† **0.9973**	0.9240	0.9592	0.9745	0.9769

One explanation for the very high accuracy in general might also be the dataset. Given that we started the exploration from location entities on DBPedia and LinkedGeoData, the overall dataset was biased towards entities from DBPedia. Hence, we intend to extend the evaluation to see if the quality of the supervised approaches remains at a comparable level, when using larger and even more diverse datasets.

6 Related Work

Previous work related to this paper can be found in three areas, each of which will be described below: (a) Extraction of geographic entities provides a starting point for our approach. The fields of (b) focused crawling on the WWW and (c) machine learning applied to Linked Data in general each share some similarities with our classification and ranking task, although differences do exist.

6.1 Extraction of Geographic Entities

Work done in the TRIDEC project [7] examined how geographic databases such as Geonames, OpenStreetMap and GooglePlaces could be used to avoid the need for error prone named entity recognition and thus increase the overall precision when geoparsing large volumes of Twitter reports for crisis mapping. This work directly compared crisis maps from Twitter with official post-disaster environment agency impact assessments, highlighting just how accurate maps based on large-scale geospatial report crowd sourcing can be. We are building on this approach within the REVEAL project and extending it by adding a Linked Data contextual lookup capability to provide better report summaries for end users and evidence for a knowledge-based trust model to improve the trust and credibility of reported data.

6.2 Focused Crawling on the WWW

The problem of prioritising (or classifying) outgoing edges of a graph without further knowledge about the linked node has been studied in the field of focused Web crawling for some time [2]. Although the approach described in this paper is not directly a crawling task, a focused crawler faces similar problems: Given a Web document, it must determine the order in which to follow outgoing links. This decision has to be

made without having seen the content of the linked document. The only information available is the current document's content as well as anchor texts and URIs of outgoing links. This scenario is similar to our proposed approach where we only know the current entity (including triples that describe it) and predicates of outgoing links.

To address this prediction problem, some focused crawlers utilize supervised machine learning classification. Chakrabarti et al. [2] and Diligenti et al. [3] present approaches that use Naive Bayes classifiers. There is however a difference in the features being used compared to our setting: While Web crawlers are restricted to textual features (like *tf-idf*-weighted vector representations of documents), our machine learning approach uses a binary predicate type feature vector that can leverage the semantic information of Linked Data. Pant and Srinivasan [10] compare different classification schemes for focused crawling and evaluate performance of Naive Bayes classifiers, SVMs and neural networks in this scenario. Micarelli and Gasparetti [**?**] give a survey of focused crawling in general and adaptive variants in particular. Ahlers and Boll [1] propose a crawler with geospacial focus that addresses the related task of retrieving Web pages containing location information. Besides the differences in features, this differs from our approach in that the crawler uses a *lookahead* to also follow pages that lead to locations indirectly. For our task we are only interested in directly linked resources.

6.3 Machine Learning over Linked Data

Machine learning approaches have been applied to Linked Data in the past for the task of predicting (or rather suggesting) additional properties that could be relevant to a resource. This problem is different from the one we address in this paper though, in that the properties of the examined resource are known in advance. In our scenario, the only information we have to make a prediction is the predicate that is used to link to the resource. Oren et al. [9] address the problem of predicting predicates with a classifier that uses *containment* and *resemblance* similarity metrics to generate a ranking of suggested predicates, in addition to a co-occurrence-based approach based on association rule mining. Nickel et al. [8] present a machine learning approach to Linked Data based on the factorisation of a sparse tensor, building on the idea "that reasoning can often be reduced to classifying the truth value of potential statements". This technique can be used to predict unknown triples for a resource, as well as for the retrieval of similar resources. Furthermore there have been efforts to use machine learning for statistical schema induction, i. e. gathering ontological knowledge from RDF datasets [12].

7 Summary and Conclusions

In this paper we addressed the task of focused exploration of Linked Open Data. We motivated the task from a concrete use case setting of exploring Linked Data entities with geo-coordinates and provide a formalisation of the task under two points of view: (a) as a classification task and (b) as a ranking task. We then present five different approaches based on the paradigms of using schema semantics, of performing statistical learning and by adapting weights from the field of Information Retrieval. In an empirical evaluation we compared the performance of all approaches and observed high

levels of accuracy when using machine learning techniques for implementing a focused exploration task.

As future work we will investigate the idea of focused exploration also for additional use cases and types of information. We will also investigate adaptive methods which adjust to new data and continue to learn while they discover new entities.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013), REVEAL (Grant agree number 610928).

References

1. Ahlers, D., Boll, S.: Adaptive geospatially focused crawling. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 445–454. ACM (2009)
2. Chakrabarti, S., Van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks 31(11), 1623–1640 (1999)
3. Diligenti, M., Coetzee, F., Lawrence, S., Giles, C.L., Gori, M., et al.: Focused crawling using context graphs. In: VLDB. pp. 527–534 (2000)
4. Gottron, T.: Of Sampling and Smoothing: Approximating Distributions over Linked Open Data. In: PROFILES'14: Proceedings of the Workshop on Dataset ProfiIling and Federated Search for Linked Data (2014)
5. Gottron, T., Gottron, C.: Perplexity of Index Models over Evolving Linked Data. In: ESWC'14: Proceedings of the Extended Semantic Web Conference. pp. 161–175 (2014)
6. Linking Open Data cloud diagram: Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/ (2011), this work is available under a CC-BY-SA license.
7. Middleton, S.E., Middleton, L., Modafferi, S.: Real-time crisis mapping of natural disasters using social media. IEEE Intelligent Systems 29(2), 9–17 (2014)
8. Nickel, M., Tresp, V., Kriegel, H.P.: Factorizing yago: Scalable machine learning for linked data. In: Proceedings of the 21st International Conference on World Wide Web. pp. 271–280. WWW '12, ACM, New York, NY, USA (2012)
9. Oren, E., Gerke, S., Decker, S.: Simple algorithms for predicate suggestions using similarity and co-occurrence. In: The Semantic Web: Research and Applications, pp. 160–174. Springer (2007)
10. Pant, G., Srinivasan, P.: Learning to crawl: Comparing classification schemes. ACM Trans. Inf. Syst. 23(4), 430–462 (2005)
11. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., Decker, S.: Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: LDOW (2010)
12. Völker, J., Niepert, M.: Statistical schema induction. In: The Semantic Web: Research and Applications, pp. 124–138. Springer (2011)