Computing Geo-Spatial Motives from Linked Data for Search-driven Applications

Andreas Both[1], Liliya Avdiyenko[1], and Christiane Lemke[1]

R & D, Unister GmbH
Barfussgaesschen 11
Leipzig (Germany)
{andreas.both|liliya.avdiyenko|christiane.lemke}@unister.de

Abstract. The Web of Data puts a vast and ever-increasing amount of information at the disposal of its users. In the era of big data, interpreting and exploiting these information is both a highly active research area and a key issue for users in industry trying to gain a competitive edge.
One current problem in industry with many potential application areas is finding a common theme for varying features by generating higher level summaries. We introduce the notion of *motives* to describe these common themes. Motives can be identified for all sorts of entities such as geo-spatial regions (e.g., "cultural regions") or holidays (e.g., "winter holidays", "activity holidays"). These motives are closer to common language and human conversations than ordinary keywords.
Since users prefer formulating their information needs using everyday language, which expresses their understanding of the world, the potential for a strong industrial impact for search applications can be derived. However, capturing the users' often vaguely formulated intentions and matching them to appropriate retrieval operations on the available knowledge bases is a challenging issue. Yet, it is an important step on the way of providing the best possible search experience to users.
This paper presents our work in progress on computing motives for geo-spatial regions. Following a long term agenda, we are evaluating the requirements for identifying such motives in large data sets. At this point, we can show that out-of-the-box machine learning methods can be used on Linked Data to train a model for computation of geo-spatial motives with good accuracy.

Keywords: knowledge discovery, knowledge extraction, information retrieval, search-driven applications, machine learning

1 Introduction

While the amount of publicly accessible and machine processable data sets in the Web of (Linked) Data grows permanently, users struggle to keep up with its growing size and complexity, and often fail to use its full potential. The question of "How can this data be made searchable?" is one of the main challenges from an industrial perspective.

One approach is to annotate all available pieces of information with a probability representing the confidence in this information, e.g., like it is done in the Google Knowledge Vault [8]. Another approach is the (manual) annotation of entities with properties based on the common understanding and knowledge which an actual user would have. This has the advantage that understandable aggregations of the data exist even if the size and complexity of the knowledge base is increasing. Here, we use *motives* to describe this level of abstraction.

A well-known methodology of structuring data is to model categories and assign them to data items of the knowledge base. For example, for structuring geo-spatial entities, a good category might be "regions where educational institutions are located". Motives express the existence of a high relevance of the given characteristics (e.g., "educational institutions"). Hence, a category in the sense of a motive might be "regions *well known* for there educational institutions", i.e., it is a common interpretation of a highly relevant set of entities. In previous work, this was very successfully applied within search-driven applications [13], in particular in the travel vertical [4]. The latter publication uses *geo-spatial motives* (i.e., motives for geo-spatial regions) which are also of focus in this paper. However, these approaches use motives that are manually annotated for each entity by experts, which does not scale when faced with large and growing knowledge bases. In this paper, we pursue a (semi-)automatic[1] process of annotating motives from user feedback using publicly available data sets.

One of the crystallization points for the Web of Data is DBpedia [2] a data set created by extracting information from Wikipedia[2]. Although Wikipedia is a global knowledge base, naturally local information have to be provided by local communities most of the time, which makes information dependent on the different cultural and personal backgrounds of the contributors.

Therefore, it can be assumed that computing a motive from DBpedia is non-trivial and at least the following challenges need to be addressed:

1. local communities are active with a different intensity, e.g., there are 1.193.557 Wikipedia articles written in Polish and just 35.154 in African[3] although a similar number of people are speaking these languages,
2. the history of entity types differ with regard to culture and administrative regions of the world,
3. (administrative) regions have varying definitions depending on the country they are in.

However, the Web of Data gives us the opportunity to use further, interlinked data sets to increase the number of available features for the considered entities. This contribution uses Natural Earth data [9] and GeoNames [6] in addition to DBpedia [7].

Additional problems appear when working with globally crowdsourced Linked Data:

[1] manual annotations are needed for the training
[2] `http://www.wikipedia.org`
[3] Fetched from `www.wikipedia.org` on 2015-04-20.

1. the interpretation of data might differ from region to region, e.g., the number of educational institutions in the United Kingdom is way higher than the one in France (following DBpedia), although the importance of higher education can be considered as equal in both countries and the number of inhabitants is close to equal,
2. the data is incomplete, i.e., it cannot be assumed that all entities of a type are captured,
3. there is no certainty property available.

Calculating motives or determining their presence is hence a challenging task, as humans, naturally, will judge calculated motives by their experience, which is influenced by many factors such as cultural background, place of residence or the education level. For this paper, the presence of motives was judged by an expert committee. In an iterative process, we used machine learning techniques to train the properties (i.e., features describing the geo-spatial motives).

The paper is organized as follows: The next section presents related work. Section 3 describes data sets used in this contribution. The idea and requirements for motive computation are shown in Section 4, while a case study is performed in Section 5. Finally, the paper is concluded in Section 6, where also the future work is described.

2 Related Work

The vision of a machine-processable *Web of Data* originating from Tim Berners-Lee has been described in its aspects and principles in [1].

While some authors point out a lack of research at the intersection between Linked Data and machine learning [3], a number of related applications and algorithms have been developed, often involving the DBpedia data set: [15] describes an extension to the popular machine learning tool RapidMiner[4], allowing to integrate linked open data with conventional data sets and transparently performing advanced data analytics, including a proof-of-concept using DBpedia data. Web search results are clustered using DBpedia background knowledge in [17]. Contextual itemsets in DBpedia are mined in [16].

Furthermore, finding abstractions and grouping entities in DBpedia has been a subject of interest close to our contribution. In [12], topic modelling based on the DBpedia graph is investigated, labeling identified topics with the most promising DBpedia concept. A higher level abstraction of DBpedia entities is generated by clustering the finely-grained categories annotated in each entity and finding a common label in the Wikipedia category tree in [18]. The authors of [18] use the notion of *domain* to describe the entities and concepts of a knowledge base with the "set of broad thematic areas, or topics, they are mostly focused on". In contrast, our work is concerned with flexibly finding higher level *characteristics* of entities without having to rely on modelled concepts or categories in the original data set.

[4] http://www.rapidminer.com

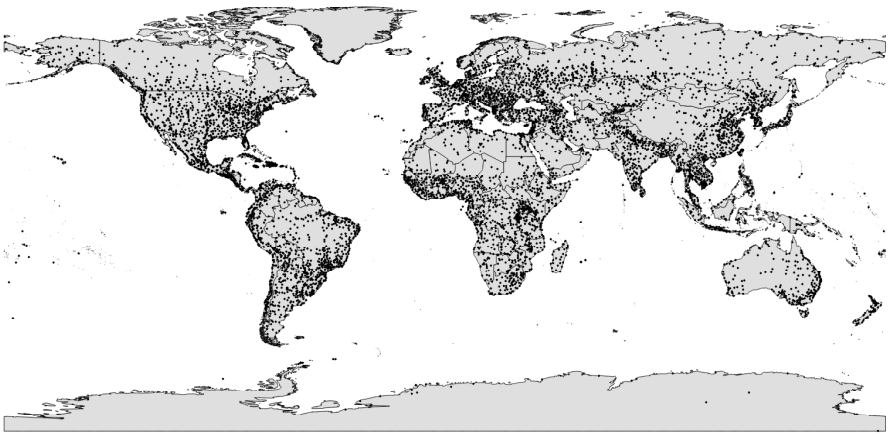

Fig. 1. World: countries and populated places.

3 Background

3.1 DBpedia

> DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia[5] and make this information available on the Web. [7]

Hence, DBpedia [14] is a structured data set computed from the information provided by Wikipedia. In this paper, the latest revision (without language-specific extension) was used. It captures the information extracted from Wikipedia in late April / early May 2014. 4218630 entities are described within the data set with 526256 entities having an annotated geo-spatial relation. Here only the property `http://www.georss.org/georss/point` is used.[6] DBpedia contains a type system facilitating an evaluation of different kind of entities.

3.2 Natural Earth

> Natural Earth is a public domain map data set (...). Featuring tightly integrated vector and raster data, with Natural Earth you can make a variety of visually pleasing, well-crafted maps (...). [9]

[5] `http://www.wikipedia.org`

[6] However, the quality of geo-spatial entities represented by the property `http://www.georss.org/georss/point` is not sufficient as for example major cities of Germany like Berlin, Cologne or Hamburg do not have this relation.

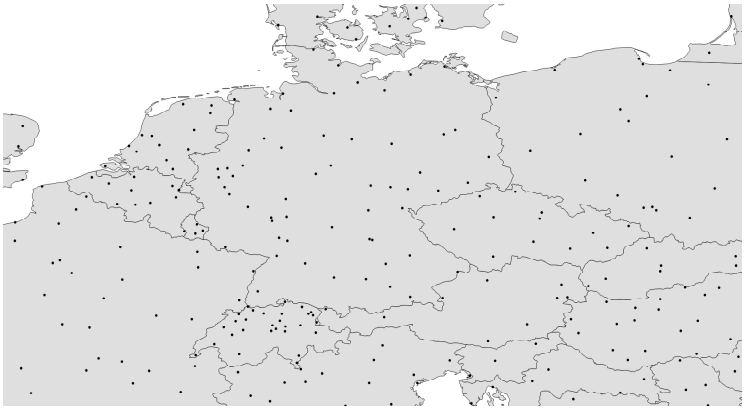

Fig. 2. Countries and populated places in a part of Central Europe.

Hence, the main purpose of Natural Earth is to provide data required for computing maps and their visualisation. However, the data is linked to GeoNames [6] and integrates available properties. The data of Natural Earth is manually selected and curated. Although there is no guarantee for one hundred percent correctness, the data set is expected to have a high quality. In this paper, we use the data sets *Admin 0 - Countries*[7] (version 3.1.0, 1:10m scale) – containing all kind of countries (called admin-0) of the world – and *Populated Places*[8] (version 3.0.0, 1:10m scale) which includes "admin-0 and many admin-1 capitals, major cities and towns, plus a sampling of smaller towns in sparsely inhabited regions". A visual representation of both data sets on one map is shown in Figure 1. As mentioned earlier, crowd-sourced data might not be consistent (c.f., Section 1, challenges). For example, considering the populated places provided by Natural Earth, very small cities are included for some countries (e.g., Gedrus, Swiss, population: 5681[9]). At the same time, major cities are not available for other regions (e.g., Halle, Germany, population: 234,107), c.f., Figure 2.

3.3 Geonames

> The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge. [6]

Besides the geo-spatial location of the entities, GeoNames also provides additional data like the population of the entities. Here, we use only the population properties that are already linked to the Natural Earth data.

[7] `http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/`

[8] `http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-populated-places/`

[9] source: GeoNames

4 Geo-spatial Motives

In this section, we will describe the properties of geo-spatial motives and collect the requirements for their computation.

> A motive can be defined as the reason for a search as well as particular conditions like how much a product should cost. [13]

Hence, motives in an information retrieval context are higher-level summaries of varying features with a common theme or topic. They are a convenient way for a user to express search queries by using vague ideas and feelings, which are more intuitive and less restrictive than classic keyword search. As an example, consider the motive "winter holidays": A user is far more likely to search for "places *ideal* for winter holidays" than for "places in the mountains with at least three ski lifts, guaranteed snow from December to March and a ski rental facilities" (which might be a common interpretation of *ideal*), even though the latter query would be easier to answer for an information retrieval system.

Accordingly, it can be assumed that a motive for a populated place exists if potential users will associate it with the related concept. We express the relation by the following definition:

Definition 1 (Geo-Spatial Motives). *A geo-spatial motive P is named as property* `urn:unister:has-geospatial-motive` *and O is an entity expressing the motive. Hence, the triple S P O expresses the relation, where S has a property of the type* `http://www.georss.org /georss/point`.

That is, the following fact might exist with regard the example from above: `dbpedia:Innsbruck urn:unister:has-geospatial-motive urn:unister:-geospatial-motive:winter-holidays`.

As the common characteristics of geo-spatial regions as perceived by users should be represented by a geo-spatial motive, we demand:

Requirement 1 (Requirements) *A motive m should be bound to a geo-spatial entity e if and only if m is a commonly associated characteristic of the considered region e.*

Hence, with regard to our available data sets (c.f., Section 3), geo-spatial DBpedia entities might be annotated with their type as motive if they are well-known within the considered country.

5 Case Study

In a case study, we will show how different data sets containing different features influence the quality of a trained model for the computation of motives. For the evaluation, we choose the *education* motive as it is present in most of the countries and can be interpreted for countries without being part of the local community. That is, populated places should be annotated with this motive if and only

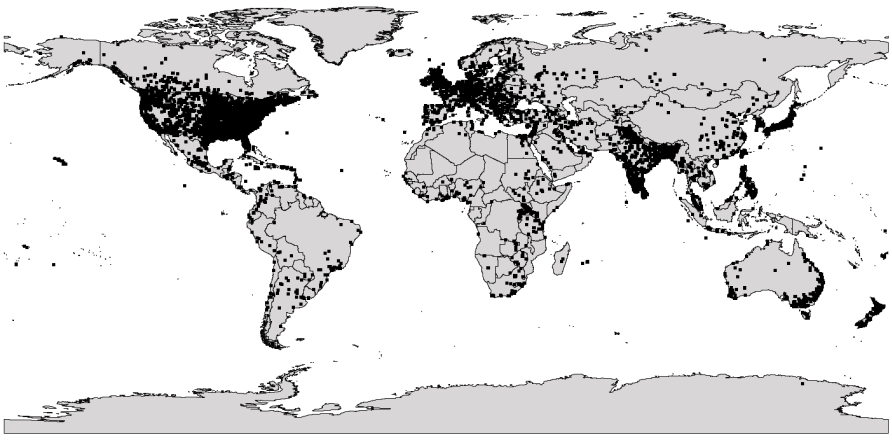

Fig. 3. World: countries and education entities (of DBpedia).

if they are well-known for their educational institutions with regard to the country they are located in. This limits the selection of relevant DBpedia entities to entities of the type `http://dbpedia.org/ontology/EducationalInstitution`, hereafter called *educational entities*. The extracted entities are shown in Figure 3 to give a visual impression. Moreover, populated places are considered if and only if there is at least one educational entity nearby[10].

5.1 Data Sets

For our experiments, we used DBpedia, Natural Earth's populated places and countries, as well as the GeoNames data set. In addition, we annotated several new features that should be useful for detecting the education relevance of populated places. First, using DBpedia, we calculated the number of educational entities within a radius $x = 10$, 25 and 50 km for every populated place (these features will be further referred as entities_nearby_$x$km). Adding information from the Natural Earth data set, which contains states and countries in which the populated places lie, it is possible to derive more sophisticated statistics:

- the maximum of the number of educational entities nearby x km over populated places in a country (referred as max_of_places_within_$x$km)
- the maximum of the number of educational entities over states in a country (referred as max_of_states)

[10] The maximal radius of the considered buffer around a populated place is defined by 50 km derived following an educated guess.

– the number of educational entities in the country (referred as e_in_country)

Figure 4 presents histograms that illustrate the number of populated places with a certain number of educational entities within the radius of 10, 25 and 50 km. As one can see, almost all places have up to $10 - 20$ entities within 10 km, whereas the distribution of places with entities within 50 km is more flat. Thus, many places have relatively few educational entities nearby, which could be an evidence for their low education relevance. However, as the Wikipedia community in some countries might be not very strong, it is also possible that not all entities are represented within the data set (we leave this for later evaluation).

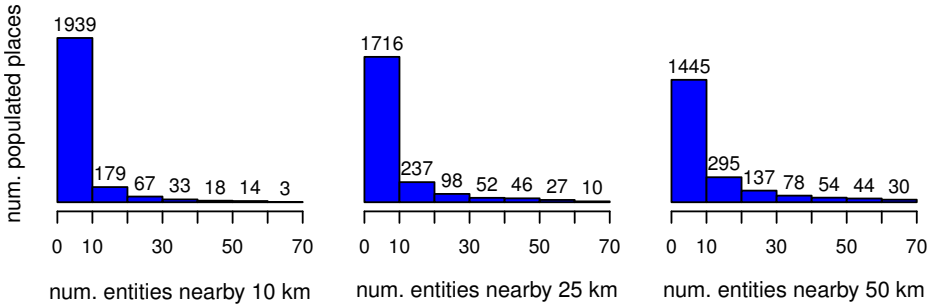

Fig. 4. Histograms illustrating a number of populated places with a certain number of educational entities within a radius of 10, 25 and 50 km.

5.2 Expert rating

The joined data set of populated places was labeled by several qualified experts. Given the name of a populated place, its state and the country, they classified the place as "highly relevant" (ranking $r = 2$), "intermediately relevant" ($r = 1$) or "irrelevant" ($r = 0$) for the education in the country of this place. Up to now, we employed three experts in the age range of $30 - 38$. Two of them have doctoral degrees in philosophy and mathematics, respectively. In addition, they have working experience in academia. The third person is currently a member of the international Masters programme in archaeology and has performed studies at several international universities. Therefore, we assume that their qualification helped them to label data as objectively as possible.

The experts were provided with a data set of 2000 places from 171 countries. They were free to rate only places which are known to them. As a result, there are 881 labeled data samples from 121 countries available for our experiments. As data classification by experts is time consuming, in the future we might infer place labels using various ranking lists of educational institutions.

5.3 Methodology

The main questions under investigation are twofold:

1. Can one use machine learning techniques on features of linked data for computing motives?
2. Is it beneficial to aggregate data from different sources?

We investigated these questions by building classifiers for motive computation and comparing their accuracy while iteratively increasing the number of data sets and features used.

In the first step, the expert ratings were merged. The resulting rating of a place is just the average of all available ratings giving a continuous value $[0, 2]$. As the number of labeled populated places is rather small, the ratings were converted into binary values. Thus, a place can have a rating of either "relevant" or "irrelevant" for the education in its country, corresponding to the initial ranges $[0, 1)$ and $[1, 2]$, respectively. In this way, the classification problem is simplified to account for the limited amount of training data. Data preprocessing and classification were done in Weka [10], an open source machine learning software.

As baseline for our experiments, we used existing and annotated DBpedia features of populated places. Of course, this data set is at least required to make a statement about the educational strength of a place. The second data set was formed by merging the extended DBpedia data set with data from Natural Earth's populated places and countries as well as features that we calculated using this data (see Section 5.1). Finally, the third data set extends the second one by GeoNames attributes of populated places. These data sets will be further referred as D1, D2 and D3, respectively. Thus, all three data sets contain 881 populated places which are described by different features. Table 1 presents features that were included in every data set, see Table 3 for the meaning of used Natural Earth and GeoNames features. These lists are not complete and contain only features that were regarded as relevant for classification by the correlation-based feature selection algorithm [11]. Note that the pop_max feature has been throttled down to the United Nations estimated metro population for the ca. 500 largest urban areas in the world[11].

As a classifier, we picked Breiman's random forest [5], which was chosen as the best classification model in terms of the F-measure and the accuracy by the cross-validation process for all data sets (see Table 5 for details). It should be noted that the choice of a classifier is not important for the present work. Our aim is rather to show that it is possible to automatically detect motives in data and that the accuracy of such detection can be improved by merging the data from several sources.

5.4 Evaluation and Discussion

Table 2 presents evaluation metrics of classifiers build on different data sets (Table 4 gives the detailed explanation of the analyzed metrics). All metrics are

[11] c.f., http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-populated-places/

Table 1. Classifier features for three data sets: DBpedia D1, DBpedia extended by Natural Earth populated places and countries D2, DBpedia extended by Natural Earth populated places, countries and GeoNames D3

D1	D2	D3
entities_nearby_10km	*all features of D1 and*	*all features of D2 and*
entities_nearby_25km	max_of_states	gn_pop
entities_nearby_50km	e_in_country	
	worldcity	
	pop_max	
	rank_max	
	max_areami	

Table 2. Evaluation metrics for different data sets: true positive rate, false positive rate, precision, F-measure, AUC (area under the ROC curve) and accuracy

	TP rate	FP rate	Precision	F-measure	AUC	Accuracy
D1	0.73	0.57	0.70	0.71	0.64	0.73
D2	0.79	0.42	0.78	0.78	0.82	0.79
D3	**0.80**	**0.41**	**0.79**	**0.79**	**0.83**	**0.80**

averaged over 200 runs. For every run, the original corresponding data set with 881 samples was randomized and divided into a training and a testing set containing 66% and 34% of the original data, respectively. Despite the fact that the classification accuracy is not perfect, in our opinion, the results are promising. One can clearly see the classifier built on D3, the data set containing DBpedia, Natural Earth and GeoNames data, achieving the best results. Moreover, enriching DBpedia data only with the features from Natural Earth improves the classification performance drastically. Though, the difference between D2 and D3 is not large, all metrics achieved on D3 are significantly better that those achieved on D2 according to the Wilcoxon signed-rank test at the p-level$= 0.05$. Adding just a single feature from GeoNames, gn_pop, it is possible to achieve the higher values of the accuracy and the F-measure, which illustrates the general classifier performance and its performance w.r.t. the positive class only, respectively. This means that using features of D3, one can classify both educationally relevant and irrelevant populated places more precisely. Thus, one obviously profits from merging various data sets while trying to automatically detect motives. Many data sets, if merged properly, provide a large pool of features that can be useful for detecting different motives.

Detailed evaluation results of the classifier trained on D3 are presented in Table 6 in the appendix. The table contains metrics of 5 runs for every class separately and their weighted average. Note that training data is highly unbalanced having 30% educationally relevant and 70% irrelevant populated places. This fact explains the difference in values of the evaluation metrics for two classes.

Note that the feature selection process, which was run on D2, selected our annotated features as relevant for classification. Moreover, it seems that infor-

mation over all distances, 10, 25 and 50 km, is relevant, which might account for differences between regions with the diverse population density. Therefore, we assume that including more features concerning local geo- and demographic statistics will help to improve the classification accuracy.

6 Conclusion and Future Work

Making knowledge bases searchable is a key challenge considering the data economics. Hence, both academics and industry have to increase their effort to make knowledge bases accessible for actual (human) users. Geo-spatial motives are one approach to enable end-users to connect their common sense of the world knowledge with the actual representation within the knowledge base.

In this paper, we presented and evaluated geo-spatial motives, an approach for a more general and flexible representation of the data available in knowledge bases. With these geo-spatial motives, we will push the available data towards better understandability and aim for better search-driven applications.

We have shown that machine learning techniques on Linked Data can be used for computing geo-spatial motives. Moreover, it seems crucial to include the knowledge of different data sets for good quality of the calculated motives. Despite the different challenges triggered by the properties of the available data sets, it was possible to achieve a promising motive detection quality within the considered case study. Therefore, we can assume that industrial search-driven applications can take advantage of knowledge bases in the Web of Data to provide a more human-friendly interaction.

However, this paper is just one step on our research agenda working towards knowledge bases that can be used and be of benefit for both experts and non-experts. In the future, we will evaluate the capabilities of different knowledge bases and generalize our approach with the aim of preventing a training for all expected motives. Instead, we will establish mechanisms that work by analogy. Of great impact might be the integration of different levels of geo-spatial regions like admin-1 entities (states, provinces) of Natural Earth[12] to increase the quality and to extend the considered motives from populated places to natural regions like the Alps (e.g., winter sports) or the Silicon Valley (e.g., IT companies). Finally, it has to be evaluated how the different personal backgrounds of users (e.g., education level, cultural background, age) is influencing their perception of motives.

Acknowledgments We thank Luise Erfurth, Stephan Schwinger, Bernd Eickmann and Kristin Mittag for their valuable support. This work has been supported by grants from the European Union's 7th Framework Programme provided for the project GeoKnow (GA no. 318159).

[12] `http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-1-states-provinces/`

References

1. Auer, S., Hellmann, S.: The web of data: Decentralized, collaborative, interlinked and interoperable. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: science, services and agents on the world wide web 7(3), 154–165 (2009)
3. Bloem, P., de Vries, G.K.: Machine learning on linked data, a position paper. Linked Data for Knowledge Discovery p. 69 (2014)
4. Both, A., Keck, M., Nguyen, V., Henkens, D., Kammer, D., Groh, R.: Get Inspired: A visual divide and conquer approach for motive-based search scenarios. In: 13th International Conference WWW/Internet (2014)
5. Breiman, L.: Random forests. In: Machine Learning. pp. 5–32 (2001)
6. geographical database, G.: http://www.geonames.org/, accessed 2015-03-15
7. DBpedia.org: http://dbpedia.org/, accessed 2015-03-15
8. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610. ACM (2014)
9. Earth, N.: http://www.naturalearthdata.com/, accessed 2015-03-15
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009), http://doi.acm.org/10.1145/1656274.1656278
11. Hall, M.A.: Correlation-based feature selection for machine learning. Tech. rep., Hamilton, New Zealand (1999)
12. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised graph-based topic labelling using dbpedia. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 465–474. ACM (2013)
13. Keck, M., Herrmann, M., Both, A., Gaertner, R., Groh, R.: Improving motive-based search. In: Streitz, N., Stephanidis, C. (eds.) Distributed, Ambient, and Pervasive Interactions, LNCS, vol. 8028, pp. 439–448. Springer Berlin Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-39351-8_48
14. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web (2014)
15. Paulheim, H., Ristoski, P., Mitichkin, E., Bizer, C.: Data mining with background knowledge from the web. RapidMiner World (2014)
16. Rabatel, J., Croitoru, M., Ienco, D., Poncelet, P.: Contextual itemset mining in dbpedia. In: 1st Workshop on Linked Data for Knowledge Discovery (LD4KD) co-located with ECML PKDD'2014: The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. vol. 1232, pp. http–ceur. CEUR (2014)
17. Schuhmacher, M., Ponzetto, S.P.: Exploiting DBpedia for web search results clustering. In: Proceedings of the 2013 workshop on Automated knowledge base construction. pp. 91–96. ACM (2013)
18. Titze, G., Bryl, V., Zirn, C., Ponzetto, S.P.: DBpedia Domains: augmenting DBpedia with domain information. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014) (2014)

A Appendix

Table 3. Natural Earth and GeoNames features used in experiments

Shortcut	Meaning
worldcity	indicates whether a city is important for the global economics
pop_max	the population of the city metropolitan area
rank_max	a population rank calculated based on pop_max
max_aremi	a metropolitan area of the city in squared miles
gn_pop	the city population according to the GeoNames

Table 4. Evaluation metrics for a binary classifier (defined in terms of positives (P), negatives (N), true positives (TP), true negatives (TN) and false positives (FP))

Metric	Definition
True positive rate (recall)	$TPR = TP/P$
False positive rate	$FPR = FP/P$
Precision	$Pr = TP/(TP + FP)$
F-measure	$F = 2 * Pr * TPR/(Pr + TPR)$
AUC	area under the curve depicting TPR plotted against FPR
Accuracy	$Acc = (TP + TN)/(P + N)$

Table 5. Evaluation metrics of several classifiers run on D1, D2 and D3

		Naive Bayes	kNN with $k = 3$	Decision Tree (J48)	SVM	Random Forest
D1	F-measure	0.67	0.70	0.70	0.67	**0.71**
	Accuracy	0.73	0.74	**0.75**	**0.75**	0.73
D2	F-measure	0.77	0.77	0.75	0.75	**0.78**
	Accuracy	0.77	0.78	0.78	**0.79**	**0.79**
D3	F-measure	0.78	0.78	0.76	0.77	**0.79**
	Accuracy	0.79	0.79	0.78	**0.80**	**0.80**

Table 6. Evaluation metrics of a classifier trained on D3

	Class	TP Rate	FP Rate	Precision	F-Measure	AUC	Accuracy
Run 1	relevant	0.514	0.097	0.627	0.565	0.842	0.809
	irrelevant	0.903	0.486	0.854	0.878	0.842	0.809
	weight. avg.	**0.809**	**0.392**	**0.799**	**0.802**	**0.842**	**0.809**
Run 2	relevant	0.513	0.117	0.6	0.553	0.815	0.789
	irrelevant	0.883	0.487	0.842	0.862	0.815	0.789
	weight. avg.	**0.789**	**0.392**	**0.780**	**0.783**	**0.815**	**0.789**
Run 3	relevant	0.494	0.086	0.672	0.569	0.826	0.803
	irrelevant	0.914	0.506	0.834	0.872	0.826	0.803
	weight. avg.	**0.802**	**0.395**	**0.791**	**0.792**	**0.826**	**0.803**
Run 4	relevant	0.467	0.107	0.593	0.522	0.804	0.786
	irrelevant	0.893	0.533	0.833	0.862	0.804	0.786
	weight. avg.	**0.785**	**0.426**	**0.773**	**0.776**	**0.804**	**0.786**
Run 5	relevant	0.488	0.082	0.684	0.569	0.849	0.803
	irrelevant	0.918	0.513	0.831	0.872	0.849	0.803
	weight. avg.	**0.802**	**0.397**	**0.791**	**0.791**	**0.849**	**0.803**
average		**0.797**	**0.400**	**0.787**	**0.789**	**0.827**	**0.798**