Popularity-Driven Ontology Ranking Using Qualitative Features

Niklas Kolbe[1(✉)], Sylvain Kubler[2], and Yves Le Traon[1]

[1] University of Luxembourg, Luxembourg, Luxembourg
{niklas.kolbe,yves.letraon}@uni.lu
[2] Université de Lorraine and CRAN, Vandœuvre-lès-Nancy, France
s.kubler@univ-lorraine.fr

Abstract. Efficient ontology reuse is a key factor in the Semantic Web to enable and enhance the interoperability of computing systems. One important aspect of ontology reuse is concerned with ranking most relevant ontologies based on a keyword query. Apart from the semantic match of query and ontology, the state-of-the-art often relies on ontologies' occurrences in the Linked Open Data (LOD) cloud to determine relevance. We observe that ontologies of some application domains, in particular those related to Web of Things (WoT), often do not appear in the underlying LOD datasets used to define ontologies' popularity, resulting in ineffective ranking scores. This motivated us to investigate – based on the problematic WoT case – whether the scope of ranking models can be extended by relying on qualitative attributes instead of an explicit popularity feature. We propose a novel approach to ontology ranking by (i) selecting a range of relevant qualitative features, (ii) proposing a popularity measure for ontologies based on scholarly data, (iii) training a ranking model that uses ontologies' popularity as prediction target for the relevance degree, and (iv) confirming its validity by testing it on independent datasets derived from the state-of-the-art. We find that qualitative features help to improve the prediction of the relevance degree in terms of popularity. We further discuss the influence of these features on the ranking model.

Keywords: Learning to rank · Ontology reuse · Web of Things · Linked vocabularies · Semantic interoperability

1 Introduction

In the Semantic Web, efficient ontology reuse is a key factor to enable and enhance the interoperability of computing systems [29]. Approaches to ontology ranking are a key component in finding and selecting the most relevant ontologies based on a query [25]. The importance of ontology reuse is also increasing in Internet of Things (IoT) environments, in which the adoption of Semantic Web technologies has received great interest [2,4]. Emerging open innovation IoT ecosystems [15] aim for the seamless discovery, access and integration of

C. Ghidini et al. (Eds.): ISWC 2019, LNCS 11778, pp. 329–346, 2019.
https://doi.org/10.1007/978-3-030-30793-6_19

heterogeneous, sensor-originated data through the Web, also referred to as the Web of Things (WoT). Efficient ontology reuse for the semantic annotation of data streams based on existing ontologies is thus a prerequisite to overcome this semantic interoperability challenge in the WoT [15]. Moreover, it enables reasoning over data and establishing linkage to existing knowledge on the Web.

Motivation. This work is motivated by the need of researchers and practitioners to discover and select the most relevant ontologies for their needs. The large number of available ontologies and the fast-paced developments in domains often make it difficult to find and select the most appropriate ontologies. For the WoT case, this is evidenced through extensive surveys in the literature [1,10,13,16]. This does not only concern ontologies with regard to sensors and sensor network setups, but further to sensor observations [13] (e.g., in the context of smart city use cases with regard to the environment, transportation, health, homes, and factories). At the core of many state-of-the-art tools that facilitate ontology reuse – such as repositories, search engines and recommender systems – lies the ranking of ontologies for a user query in the form of keywords.

Importance of Popularity. Fundamental ontology reuse strategies rely on ontologies' *popularity*, which is typically understood as the measure of how often an ontology is used to model data in the Linked Open Data (LOD) cloud [27]. While rankings foremost take into account the semantic match of query and ontologies in the collection, current state-of-the-art tools such as Linked Open Vocabularies (LOV) [32], TermPicker [28], and vocab.cc [30] further incorporate such a popularity measure in their ranking model. This is crucial because it reflects the community's consensus on ontologies' relevance, instead of solely relying on how well ontologies semantically match the query. Thus, the approach of computing the popularity measure has an important influence on the performance of the ranking model.

Problem Statement. We find that the approach to derive popularity from LOD datasets, as computed in many state-of-the-art tools, can be problematic for ontologies of some domains. We illustrate this problem in Fig. 1, which shows the number of ontologies contained in the well-known LOV platform that have never been reused in LOD datasets[1]. In total, only ∼35% of the ontologies in the repository have been reused. We identify particular critical domains with no reuse in any LOD dataset for any ontology in the collection, namely: Services, Industry, IoT, Transport, and Health. We consider all these domains highly relevant to WoT application domains (e.g., smart mobility, smart health care, industry 4.0), which thus forms our motivating case to investigate qualitative ontology ranking from this perspective. From a more general viewpoint, this case highlights the problem that the likeliness of missing relevant information to

[1] Extracted from the LOV SPARQL endpoint: https://lov.linkeddata.es/dataset/lov/sparql – accessed 03/2019.

explicitly determine popularity for all ontologies in a collection is high, leading to the computation of ineffective popularity scores in the ranking model.

Contributions. This research contributes to the extension of scope and effectiveness of popularity-driven ontology ranking models, aiming to make these models less dependent on the underlying popularity measure, such as the selection of LOD datasets (and the way these datasets are assembled). In this respect, we investigate whether the relevance degree in terms of popularity can be predicted with qualitative properties of the ontology instead of relying on an explicit popularity feature as it is common in the state-of-the-art. We perform this study (based on the problematic WoT case) by learning a ranking model that uses the popularity as relevance degree for the prediction target. This approach to ontology ranking results in fairer scores for ontologies that were developed for use cases other than LOD publication, such as semantic sensor data annotation and the development of context-aware applications. In general, obtaining relevance labels for learning to rank is perceived as a major challenge and a costly

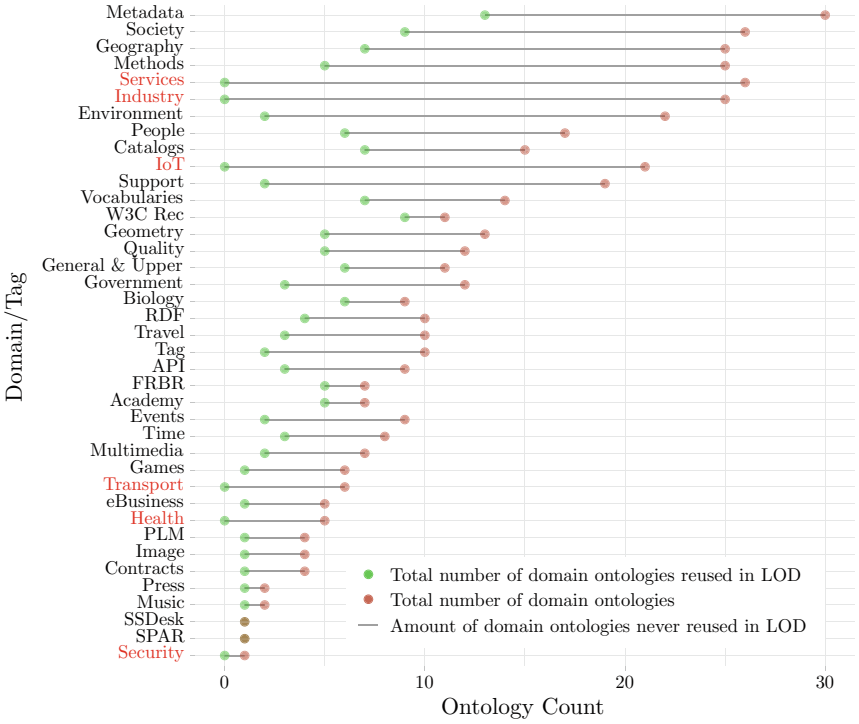

Fig. 1. Count of ontologies per category in the LOV repository that were never reused in LOD datasets, which is often used as underlying popularity measure in state-of-the-art rankings. It shows that this score is inefficient for many domains related to WoT applications, for which none of the ontologies appear in any LOD dataset.

Table 1. Notation.

Var.	Meaning	Function	Meaning
q	Keyword query	$\Phi(q, o, R)$	Relevance feature extractor
q_i	i^{th} term of query q	$\Phi(o, R)$	Importance feature extractor
o	Ontology	$\mathrm{TF}(q_i, o, R)$	Term frequency
R	Ontology repository	$\mathrm{IDF}(q_i, R)$	Inverse document frequency
M_{w2v}	Word2Vec vector space	$\mathrm{coord}(q, o)$	Scoring for number of q_i matches
D_{WN}	WordNet dictionary	$\mathrm{queryNorm}(q)$	Normalization factor
w_i	i^{th} word in collection w	$\mathrm{propertyBoost}(q_i, R)$	Boost based on matched property
Φ_i	i^{th} feature	$\mathrm{cosineDistance}(q, w_i, M_{w2v})$	Similarity of query and word
l	Relevance judgment	$\mathrm{sense}(q, D_{WN})$	Senses of query (WordNet)
π_l	Total order	$\mathrm{synonym}(q, D_{WN})$	Synonyms of query (WordNet)

process [17]. We propose a popularity measure for ontologies of WoT domains that relies on scholarly data (i.e., the citation history of ontologies' associated scientific publication) to determine relevance degrees in terms of popularity. This approach overcomes limitations of existing approaches, and we ensure that this measure approximates popularity in terms of reuses by evaluating the model on state-of-the-art rankings.

The remainder of this paper is structured as follows. The background and related work for ontology ranking are presented in Sect. 2. Section 3 defines the key ranking features and introduces the approach to relevance mining from scholarly data. The experiments, data collection and results are presented in Sect. 4. The findings are further discussed in Sect. 5; the conclusion follows.

2 Background and Related Work

This section introduces the background regarding ontology ranking, learning to rank and related work. The notation in this paper is summarized in Table 1.

2.1 Ontology Ranking and Learning to Rank

Approaches to ontology ranking adopt conventional ranking techniques and models from information retrieval, which can be categorized as follows [17]: *relevance ranking models* aim to rank an ontology o from a repository R based on their relevance to a query q, i.e., in the form of $\Phi(q, o)$ or $\Phi(q, o, R)$. These include well-known approaches (e.g., TF-IDF [26], BM25 [24]) and further ontology-specific approaches such as centrality of matched concepts in the ontology graph [8]. On the other hand, one can find *importance ranking models* that rank ontologies independently from the query, i.e., in the form of $\Phi(o)$ or $\Phi(o, R)$. Models that compute scores based on the quality of ontologies in a collection belong to this category. Well known approaches include PageRank [22]; ontology-specific

approaches consider qualitative metrics such as ontologies' popularity, availability, interlinkage to other ontologies, etc. [15]. Some ontology ranking models have been studied in [7].

In most practical settings various of the previous introduced scoring functions Φ are combined to form a better performing ranking model $h(q, o, R)$. Learning-to-rank approaches allow to automatically tune the parameters when combining different ranking models by employing supervised machine learning algorithms [17]. The parameters are derived based on the correlations of features (i.e., relevance and importance scores) and a corresponding label that determines how relevant an ontology for a query is. Therefore, in order to obtain a training set for learning to rank of ontologies, one requires a ground truth that provides information about which ontologies o in a collection R are more relevant than others for a certain query q. Such a ground truth is obtained by (i) selecting a set of queries with a set of relevant ontologies per query, and by (ii) assigning relevance judgments l to each query-ontology pair. Obtaining a ground truth is a difficult task and annotating data with human assessors is costly [17]. Thus, several approaches are employed to automatically mine a ground truth by deriving labels from sources such as user click logs of existing search engines and exploiting usage patterns in LOD datasets. However, such approaches also have their limitations, e.g., using user click logs requires access to back-ends of existing search engines with a large user base, which are usually closed systems.

2.2 Related Work

Learning-to-rank techniques have been previously applied to build ontology ranking models. The CBRBench ground truth [6] was gathered through human labeling based on how well ontology terms meet their definition in a dictionary, comprising ten queries with a total of 819 relevance judgments. CBRBench was used to learn a ranking model in DWRank [8]. Termpicker [28] proposes a ground truth derived from LOD datasets and a ranking model that relies on popularity features, offering ontology term recommendations upon a query in form of triple patterns. In CARRank [33], a ground truth was obtained through human labeling for evaluation purposes, resulting in ∼400 query-term relevance judgments. Our work differs from these efforts, as we aim to rank ontologies instead of terms. Further, we aim to propose a ranking that uses popularity as a target instead of a feature, which is not captured in existing ground truths.

Ontology ranking models have been integrated in tools that help users to find and select relevant ontologies according to their need, such as the previously mentioned LOV platform [32], TermPicker [28], and vocab.cc [30]. Such tools have been previously surveyed in the literature, as in [15]. Ontology reuse has been studied from more holistic viewpoints, such as methodological guidelines [11] and choosing ontologies from a set of candidates [14]. This study contributes to ontology ranking with the overall aim to support the ontology reuse task and to improve related tools.

Ontology catalogs exist that aim at the collection and curation of ontologies related to WoT applications. The respective tools provide extensive lists of

Table 2. Overview of selected ranking features.

Category	Feature		Description
Relevance	Φ_1	Lucene	A Lucene match with property boost
	Φ_2	Word2Vec	Score based on closely related words of the query
	Φ_3	WordNet	Score based on senses and synonyms of the query
Importance	Φ_4	Availability	Whether the ontology is accessible at its URI
	Φ_5	Believability	Whether provenance information is provided
	Φ_6	Understandability	To which degree terms are labelled and commented
	Φ_7	Interlinking	To which degree the ontology refers to external terms
	Φ_8	PageRank	The importance derived through *owl:imports* statements
	Φ_9	Consistency	Whether a reasoner does not detect inconsistencies
	Φ_{10}	Richness (Width)	The size of the ontology in terms of width
	Φ_{11}	Richness (Depth)	The size of the ontology in terms of depth

ontologies and respective metadata, such as classifications, characteristics (e.g., ontology language), and background information. We are aware of three related projects: LOV4IoT[2] [12], the Smart City Ontology Catalogue[3] [23], and the Smart City Artifacts Web Portal[4] [3], which maintain an expert selection of respectively 499, 70, and 124 ontologies[5]. Whereas these projects provide valuable ontology collections for WoT application domains, to the best of our knowledge, no ranking mechanism that effectively considers these ontologies' popularity exists. We base our experiments on the collection of the LOV4IoT catalog as it contains the largest number of ontologies and more extensive metadata about the collection.

3 Ranking Features and Relevance Mining

This section presents the selected ranking features that are considered to constitute our proposed model as well as our approach to derive relevance labels for ontologies of WoT application domains. The selection of ranking features is based on comprehensive studies in the literature on ontology ranking and quality [15,34]. We include all attributes identified in survey [15] except for subjective features and those that only concern term ranking, not ontology ranking. Table 2 provides an overview of the selected features. Our interpretation of these features, as presented in the following, is guided by the review presented in [34].

3.1 Relevance Features

Relevance features aim to determine most suitable matches for a query and an ontology corpus, for which the following features are selected:

[2] http://lov4iot.appspot.com/.
[3] http://smartcity.linkeddata.es/.
[4] http://opensensingcity.emse.fr/scans/ontologies.
[5] Accessed 03/2019.

Lucene Match (Φ_1). Our fundamental feature to find relevant ontologies based on keywords is a Lucene match [19]. As argued in [32], ontologies are structured documents and more meaningful matches should be given a higher score. We adopt the approach of [32] and apply a property boost to the lucene match that aims at rewarding more important matches, such as local names, primary labels (e.g., *rdfs:label*), and secondary labels (e.g., *rdfs:comment*). The definition of the Lucene score is given in Eq. 1.

$$\begin{aligned}\text{Lucene}(q, o, R) = \ & \text{coord}(q, o) \cdot \text{queryNorm}(q) \cdot \\ & \sum_{i=1}^{n} \left(\text{TF}(q_i, o, R) \cdot \text{IDF}(q_i, R)^2 \cdot \text{propertyBoost}(q_i, R) \right)\end{aligned} \tag{1}$$

Word2Vec (Φ_2). Word2Vec [20] trains a neural network to predict the surroundings of a word. We employ this approach to find closely related words of the input search terms and compute a score based on the cosine distance and the lucene match. The respective matching score is given in Eq. 2.

$$\text{Word2VecMatch}(q, o, R) = \sum_{w_i \in \text{cosineDistance}(q, M_{w2v})} \text{cosineDistance}(q, w_i, M_{w2v}) \cdot \text{Lucene}(w_i, o, R) \tag{2}$$

WordNet (Φ_3). WordNet [21] is a lexical database in English. We use this source to find senses and synonyms of the keyword input and compute a score for these words based on the Lucene search, as given in Eq. 3.

$$\text{WordNetMatch}(q, o, R) = \sum_{\substack{w_i \in \text{sense}(q, D_{\text{WordNet}}) \ \cup \\ w_i \in \text{synonym}(q, D_{\text{WordNet}})}} \text{Lucene}(w_i, o, R) \tag{3}$$

3.2 Importance Features

Importance features aim to assign a score to an ontology within a collection independently from the query. The selected features that represent the ontologies' quality are defined as follows:

Availability (Φ_4). The availability indicates whether ontology o can be accessed at its indicated URI. We derive this feature as given in Eq. 4.

$$\text{Availability}(o) = \begin{cases} 1, & \text{if httpResponseCode}(\text{URI}(o)) = 200 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Believability (Φ_5). The believability of a published ontology increases with the presence of provenance data (e.g., specification of authors and descriptions), and is computed based on DCMI metadata terms[6], as given in Eq. 5.

$$\text{Believability}(o) = \begin{cases} 1, & \text{if } \{\text{URI(o)}\ dc:creator\ ?c\} \cup \\ & \quad \{\text{URI(o)}\ dc:description\ ?d\} \neq \varnothing \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

[6] http://purl.org/dc/terms/.

Understandability (Φ_6). The better a ontology is documented, the easier it is to reuse it. We measure the understandability of an ontology by computing how many of all defined terms in ontology o are labelled and commented.

$$\text{Understandability}(o) = \frac{|\,\text{labelledTerms}(o)|}{|\,\text{definedTerms}(o)|} + \frac{|\,\text{commentedTerms}(o)|}{|\,\text{definedTerms}(o)|} \tag{6}$$

Interlinking (Φ_7). Ontologies foster interoperability by establishing links to previously defined terms. Thus, we count the outlinks found in an ontology as formalized in Eq. 7.

$$\text{Interlinking}(o) = |\,\text{outlinks}(o)| \tag{7}$$

PageRank (Φ_8). PageRank [22] is an algorithm that helps to compute the importance of ontologies based on how often they have been referred to by others (i.e., inlinks). We compute the PageRank score based on *owl:imports* statements, as given in Eq. 8.

$$\text{PageRank}(o_i, R) = \frac{1-d}{|R|} + \sum_{o_j \in \text{importedBy}(o_i)} \frac{\text{PageRank}(o_j, R)}{|\,\text{imports}(o_j)|} \tag{8}$$

Consistency (Φ_9). Ontologies are expected to be logically consistent, which can be derived through OWL reasoners. We compute the consistency feature as given in Eq. 9.

$$\text{Consistency}(o) = \begin{cases} 1, & \text{if } \{\text{inconsistencies}(o)\} = \varnothing \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Richness (Φ_{10} & Φ_{11}). We further consider the size of the ontology in the form of its width (see Eq. 10) and depth (see Eq. 11).

$$\text{Width}(o) = |\,\text{typeStatements}(o)| \tag{10}$$

$$\text{Depth}(o) = |\,\text{subClassOfStatements}(o)| + |\,\text{subPropertyOfStatements}(o)| \tag{11}$$

3.3 Relevance Mining Approach

Learning to rank is a supervised machine learning approach that requires relevance labels for query-ontology pairs. We propose to derive a popularity measure based on corresponding scientific publications associated with an ontology. We are inspired to follow this approach as a large number of ontologies for WoT application domains emerge from research projects, as evidenced in [1,10,13,16]. Furthermore, it overcomes several limitations of other approaches: (i) as previously discussed, LOD does not provide a reliable source for ontology reuse in WoT application domains; (ii) deriving relevance through user click logs requires

access to closed back-ends of existing ontology search engines with a large user base; (iii) human labeling is costly and, unlike mining relevance from scholarly data, does not come with the benefit of being reproducible.

Our popularity score is based on two measures; (i) $citationsPerYear(o)$: citations per year are counted and divided by the number of ontologies described in the same publication to represent the overall impact of the ontology; and (ii) the $linearTrend(o)$: a linear regression of the citation history to reward positively trending ontologies combining the intercept and the slope of the linear model. The final relevance score, as given in Eq. 12, is the mean of both min-max normalized measures and used to derive the total order π_l for the set of ontologies associated with a query, for which an ontology with a higher popularity score is more relevant than another, i.e., $l_a \succ l_b$ if $popularity(o_a) > popularity(o_b)$.

$$\text{popularity}(o) = \frac{\text{citationsPerYear}(o) + \text{linearTrend}(o)}{2} \tag{12}$$

A ground truth mining process is always assumed to contain bias and noise: for relevance mining from scholarly data, all self-citations are subtracted from the citation history, and incomplete years are not considered (i.e., citations of the current year and of the year of publication). Although the citation history is often used to measure a study's impact, the associated reason for the citation remains unknown, which is a potential threat to the validity of our popularity scores. We assume that the proposed measure reflects the overall ontologies' relevance for the scientific community (e.g., we assume that for outdated ontologies the citation count will decline and the score is penalized accordingly through the linear trend). In the following experiments, the proposed ranking model is tested on completely independent datasets to evaluate whether our training data is accurate and the assumptions hold.

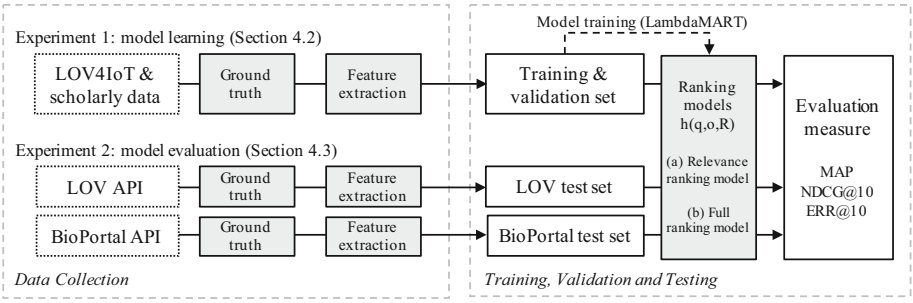

Fig. 2. Experiment overview.

4 Experiments

This section presents the experiments following the learning-to-rank approach to build a ranking model with qualitative properties of the ontologies to predict the relevance degree. An overview of the following experiments is illustrated in Fig. 2, whose aims are twofold; (i) to investigate whether qualitative features in the ranking model help to improve the ranking performance with regard to the relevance degree, and (ii) to confirm the validity of the results by testing the model on data sets derived from state-of-the-art ontology rankings.

4.1 Experiment Design

The design choices to learn and evaluate the ranking model are as follows:

Learning Algorithm: various learning-to-rank algorithms were proposed by the machine learning community. The ranking model is trained using the list-wise LambdaMART algorithm which has successfully been applied for real-world ranking problems [5] and has also been previously selected in related work for ontology ranking [8]. We rely on the LambdaMART implementation of the RankLib[7] library.

Evaluation Metrics: the performance of the ranking model is validated and tested based on the Mean Average Precision (MAP) [17], Normalized Discounted Cumulative Gain (NDCG@k) [17] and the Expected Reciprocal Rank (ERR@k) [9], considering the first ten elements ($k = 10$). A unified point-wise scale for relevance labels is required for some evaluation metrics, so popularity scores of query-ontology pairs are mapped to a scale of 0–4 for the experiments. While MAP is only a binary measure (i.e., 0: considered not relevant, 1–4: considered equally relevant), the NDCG@k and ERR@k scores do consider the multi-valued relevance labels (i.e., these metrics consider how well the ranking model matches the relevance degree 0–4). Whereas NDCG@k only depends on the position in the ranking, ERR@k discounts the results appearing after relevant ones, which supposedly better reflects user behavior of search engines [9]. The ranking model is trained by optimizing the ERR@10 score using 10-fold cross validation, meaning that the training data is randomly partitioned into ten equal sized subsamples. Iteratively, nine of these folds are used for training and the remaining one for validation.

Feature Sets: the training dataset is prepared by extracting the feature vectors for each query-ontology pair as introduced in Sect. 3. We rely on the Lucene search engine of the Stardog[8] triple store, the openllet[9] OWL reasoner to infer consistency and the GloVe word vector model[10] to compute the Word2Vec feature.

[7] https://sourceforge.net/p/lemur/wiki/RankLib/.
[8] https://www.stardog.com/.
[9] https://github.com/Galigator/openllet.
[10] https://github.com/stanfordnlp/GloVe.

4.2 Ranking Model Training and Validation

In the first experiment we train and validate the ranking model, as presented in the following.

Data Collection: the data for training and validation is collected from the LOV4IoT catalog[11]. 455 ontology files related to WoT applications could be downloaded through the catalog (each file being treated as a separate ontology). Only 433 files were syntactically correct and stored as named graphs in a local triple store. We derive training examples by using the available classification labels from the LOV4IoT catalog as queries (i.e., ontologies' domain[12] and described sensor devices[13]), and consider the correspondingly tagged ontologies as relevant. As previously motivated, we rely on scholarly data to derive degrees of relevance. From the initial collection, 395 ontologies could be assigned to 125 different scientific publications based on the LOV4IoT metadata. This collection resulted in $1.1M$ triples with $133K$ distinct terms and forms the ontology repository for the experiments. The citation history from Google Scholar of the assigned publications is used to derive a relevance score for the ontologies based on the approach presented in Sect. 3. The resulting scores are mapped to relevance labels 1–4 by dividing the range of the highest and lowest popularity score for each query into four equal-sized intervals, and a random set of irrelevant ontologies is added with the relevance label 0. The resulting ground truth contains 1028 query-ontology relevance judgments with 25 different queries, for which the previously introduced ranking features are extracted to finalize the training set.

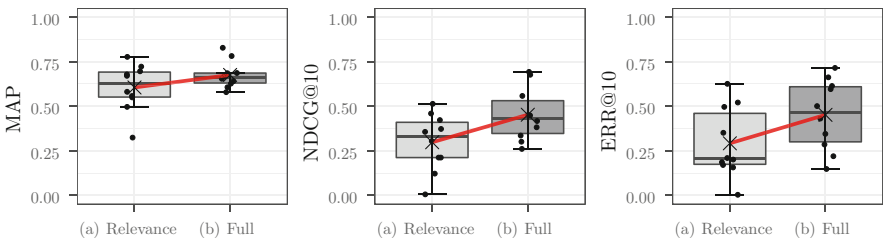

Fig. 3. Comparison of trained models with regard to MAP, NDCG@10 and ERR@10 on the validation set, for model (a) using only relevance features (Φ_1–Φ_3) and model (b) using further the importance features (Φ_1–Φ_{11}). The red lines indicate the difference of the respective metric's mean between the two models.

[11] http://lov4iot.appspot.com/.
[12] Denoted by <http://sensormeasurement.appspot.com/m3#hasContext>.
[13] Denoted by <http://sensormeasurement.appspot.com/m3#hasM2MDevice>.

Experiment and Results: the first experiment aims at investigating whether the selected qualitative importance features improve the ranking performance with regard to the relevance degree. Thus, we first train and validate a model only based on relevance features, and use this as a baseline to evaluate the performance of a model that further considers the importance features. The results are summarized in Fig. 3, showing the performance of two ranking models: the relevance model (a) is only trained with the relevance features (Φ_1–Φ_3), whereas the full model (b) also includes the importance features (Φ_1–Φ_{11}).

The results show that the trained ranking models appear to appropriately rank ontologies with regard to their relevance. We observe that the addition of qualitative features only has a small impact on the MAP score, but significantly improves the NDCG@10 and ERR@10 scores. This behavior is expected, as MAP effectively only measures the semantic match of query and relevant ontologies, whereas the qualitative features aim at ranking relevant ontologies according to their relevance degree. NDCG@10 and ERR@10 both reflect this degree, as they take into account multi-valued relevance labels. We thus conclude that qualitative features helped to improve the ranking with regard to the popularity-based relevance degree captured in the ground truth. Subsequently, this implies that the proposed approach can extend the scope of state-of-the-art rankings, by improving the performance for domains in which ontologies were never reused in LOD datasets. In such cases, the explicit popularity feature always results in the same score for all ontologies (i.e., zero) and effective ranking is only based on relevance (i.e., corresponding to model (a)). The presented approach in contrast predicts the popularity based on the qualitative features (i.e., corresponding to model (b)), even when no explicit information of popularity or reuse is present.

4.3 Ranking Model Evaluation and Comparison

The second experiment aims at evaluating and comparing the model with independent datasets derived from state-of-the-art rankings. We do this in order to ensure that our assumptions for the ground truth, as introduced in Sect. 3, hold and to confirm whether the findings from the first experiments are valid. Due to the lack of existing benchmarks and implementations of ranking models proposed in the literature, we derive test sets from state-of-the-art tools which must: (i) provide an open API that returns the computed ranking score of the top-ranked ontologies for a query; (ii) make the underlying ontology collection available for download; and (iii) incorporate a popularity measure in their ranking model. We choose to compare the proposed ranking model to approaches from two different domains that fulfill these requirements: the LOV repository [32], which measures popularity based on LOD occurrences (by excluding the problematic domains without any reuse in LOD for the test sets); and the NCBO recommender 2.0 of the BioPortal [18], which ranks biomedical ontologies and covers ontology's popularity in its notion of acceptance, derived by the number of other curated repositories that also keep an ontology in its collection.

Data Collection: we create the test sets based on the LOV REST API[14] and the BioPortal REST API[15]. For each platform, we (i) derive a set of test queries by extracting nouns and verbs from names and descriptions of all ontologies in the respective repository, (ii) use each test query to retrieve the ranking from the respective API that forms the ground truth, (iii) use the same strategy as for the training data to map the ranking scores to a scale of 1–4 and add a random set of irrelevant ontologies with a relevance of 0, and, lastly, (iv) complete the test set by extracting the features for all query-ontology pairs from a local triple store that contains the respective ontology collection. For the LOV test set we only consider domains with at least five ontologies that have been reused in LOD datasets, in order to ensure that the derived ground truth sufficiently reflects the ontologies' popularity (see Fig. 1). This process resulted in test datasets with 2998 (LOV) and 4313 (BioPortal) query-ontology relevance scores.

Experiment and Results: in the second experiment we test both, the validated relevance model (a) and the full model (b), from the first experiment on the newly derived datasets. The results are illustrated in Fig. 4, showing the comparison of the performance for the LOV and BioPortal test set, as well as the mean performance of the full ranking model from the first experiment (indicated by the dashed lines).

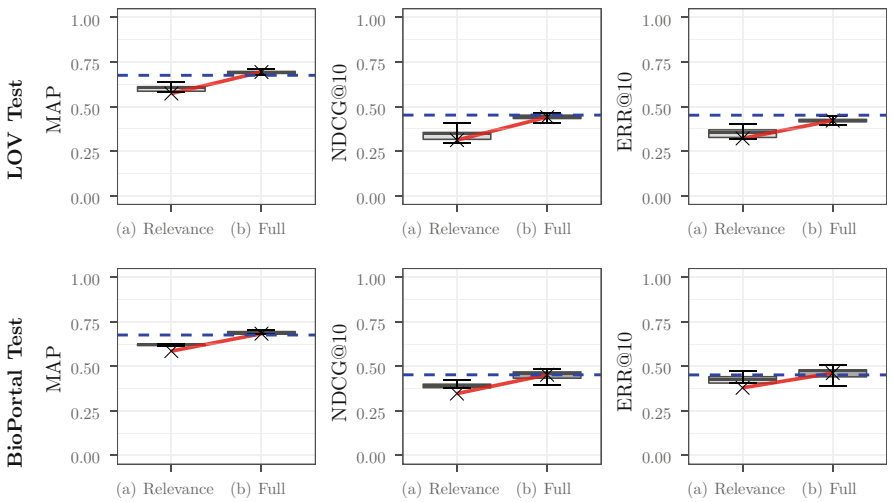

Fig. 4. Comparison of the validated ranking models from the first experiment with the LOV and BioPortal rankings. The dashed lines indicate the mean performance of the full model on the 10 fold validation sets, showing that the model performs similarly well on the test datasets. The red lines indicate the difference of the respective metric's mean between the two models. (Color figure online)

[14] https://lov.linkeddata.es/dataset/lov/api.
[15] http://data.bioontology.org/documentation.

The experiment results lead to two important conclusions. First, it shows that the learned models behave reasonably well on these completely independent datasets, evidenced by the similar performance compared to the first experiment. This confirms that the underlying ground truth to train our model is valid and, subsequently, implies that the citation history of ontologies in WoT domains is a fair approximation of their popularity. Secondly, we observe a similar behavior of the relevance and the full ranking model as in the first experiment, for which the full model improves the ranking in terms of relevance degree. Albeit the improvement on test sets is lower as in the previous experiment, it shows the same trend and thus validates our previous conclusion that the selected qualitative features help to predict the popularity-driven relevance degree of ontologies. The experimental results are further analyzed and discussed in the following.

5 Discussion

Experiment Summary. This study reveals that the prediction of ontologies' relevance for a query in terms of popularity can be improved with qualitative features. This confirms the hypothesis of a correlation between ontologies' popularity and its quality, based on the intuition that ontologies with better quality are more likely to be reused than others of the same domain. The presented approach extends the scope and applicability of the ranking model, as it is not dependent on measures of LOD occurrences. As motivated previously, this approach gives a fairer score to ontologies that are not engineered for LOD publication purposes, such as WoT application domains, and furthermore also for newly proposed ontologies without any reuses that are well-defined.

Influence of Qualitative Attributes. The LambdaMART algorithm applied in the experiments creates an ensemble of regression trees which can be further

Table 3. Full model feature frequencies averaged over all folds.

Category	Feature		Avg. freq.
Relevance	Φ_1	Lucene	1056.9
	Φ_2	Word2Vec	680.0
	Φ_3	WordNet	1375.5
Importance	Φ_4	Availability	697.7
	Φ_5	Believability	55.8
	Φ_6	Understandability	1237.9
	Φ_7	Interlinking	634.8
	Φ_8	PageRank	1302.1
	Φ_9	Consistency	777.7
	Φ_{10}	Richness (Width)	535.1
	Φ_{11}	Richness (Depth)	646.5

analyzed to better understand the model and its consequences. One way to infer the importance of each feature on the ranking model is the frequency it was used for classification of the training examples. We use these counts to discuss the model's implications and directions for future research. Table 3 reports the results for feature frequency. We derive the following insights based on the feature statistics, albeit detailed experimentation would be required to confirm them. One interesting observation is that the feature believability (Φ_5) barely contributes to the model and would be the first candidate to be replaced with another feature. This is surprising, as other approaches fundamentally rely on provenance information such as ontologies' authorship to compute the ranking [31]. Other observations include that an ontology's incoming links (Φ_8) appear to have much more significance than outgoing links (Φ_7). This is intuitive, as being imported by another ontology often requires the ontology to be considered relevant by ontology engineers other than the original authors. In addition, it can be observed that features that solely reflect the internal graph structure (Φ_{10} and Φ_{11}) are less often used by the model than more expressive qualitative scores such as understandability (Φ_6), consistency (Φ_9) and availability (Φ_4).

Implications of Proposed Ranking Approach. The experimental results of this study show that the proposed approach is promising to extend the scope of ontology ranking models. As evidenced through the experiments, this approach can also be adopted for other domains and we expect a model trained on domain-specific ontologies to perform better. This encourages further experimentation with more quality attributes, new interpretations of them, and with training sets from other domains in order to confirm the findings and achieve the development of better performing ranking models. The quality of learning-to-rank approaches also highly depends on the size of the training data. We expect future research to provide larger benchmarks that allow for the study of more complex models and better comparisons of ranking approaches, such as ground truths derived from user click logs of existing search engines. In a broader context, this approach to ranking could also encourage ontology engineers to put even more emphasis on qualitative traits of proposed ontologies in order to increase exposure and reuse in applications. Albeit the extraction of qualitative features can be computationally very expensive, these scores are independent from the user query and can be pre-computed. Thus, the lookup of these scores and re-ranking of relevant ontologies only has a minor impact on the run-time performance compared to the complexity of the semantic similarity search in the entire ontology corpus.

Novel Ontology Ranking Model for the WoT. To the best of our knowledge, the proposed full ranking model is the first that effectively considers popularity for ontologies in WoT application domains. We thus conclude that the proposed full ranking model contributes to ontology selection for these domains in the scope of open IoT ecosystems, e.g., for ontology collections such as the LOV4IoT catalog. The ranking model can be integrated in more complex user interfaces and combined with various other selection criteria in IoT domains, that, e.g., further consider important standardization efforts.

Limitations. A potential threat to validity of this study's experimental findings is the ground truth derived through popularity measures from scholarly data. While it is a common approach to use implicit user feedback as relevance score (such as user clicks), using citations arguably is a more ambiguous measure. Yet, as previously mentioned, this approach overcomes limitations of alternatives and our evaluation showed a reasonable performance. We conclude that further experimentation is required in order to confirm whether similar observations can be made for other domains than WoT, by using training examples with a relevance score derived from other popularity measures. From an ontology reuse perspective, this study is limited as it only considers ranking of single ontologies. However, practitioners often search for terms (e.g., as offered by LOV [32]) or combinations of ontologies (e.g., as offered by NCBO 2.0 [18]).

Resource Availability. The derived datasets, source files to replicate the experiments, as well as more detailed results of the ranking models are available online[16], and may be used for future experiments and comparison studies.

6 Conclusion

In this paper, we show that the prediction of ontologies' relevance in terms of popularity can be improved with qualitative features in the ranking model, making the model independent from explicit computed popularity metrics such as LOD occurrences. Moreover, we present a ranking model that effectively ranks ontologies of WoT domains with respect to their popularity. We show that the proposed model performs similarly well on test set derived from rankings of state-of-the-art tools, which is encouraging to adopt the presented approach also in other domains. Lastly, we discuss the importance of the qualitative features on the overall performance of the ranking model. The proposed model can be integrated in ontology selection mechanisms for practitioners and researchers in WoT use cases and thus contributes to establish semantic interoperability in emerging large-scale IoT ecosystems.

Acknowledgements. The research leading to this publication is supported by the EU's H2020 Research and Innovation program under grant agreement № 688203 – bIoTope.

References

1. Andročec, D., Novak, M., Oreški, D.: Using semantic web for internet of things interoperability: a systematic review. Int. J. Semant. Web Inf. Syst. (IJSWIS) **14**(4), 147–171 (2018). https://doi.org/10.4018/IJSWIS.2018100108
2. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010). https://doi.org/10.1016/j.comnet.2010.05.010

[16] Supplemental material: https://tinyurl.com/y64sa6le.

3. Bakerally, N., Boissier, O., Zimmermann, A.: Smart city artifacts web portal. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9989, pp. 172–177. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47602-5_34
4. Barnaghi, P., Wang, W., Henson, C., Taylor, K.: Semantics for the internet of things: early progress and back to the future. Int. J. Semant. Web Inf. Syst. (IJSWIS) **8**(1), 1–21 (2012). https://doi.org/10.4018/jswis.2012010101
5. Burges, C.J.: From ranknet to lambdarank to lambdamart: an overview. Learning **11**(23–581), 81 (2010)
6. Butt, A.S.: Ontology search: finding the right ontologies on the web. In: Proceedings of the 24th International Conference on World Wide Web, pp. 487–491. ACM (2015). https://doi.org/10.1145/2740908.2741753
7. Butt, A.S., Haller, A., Xie, L.: Ontology search: an empirical evaluation. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8797, pp. 130–147. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11915-1_9
8. Butt, A.S., Haller, A., Xie, L.: DWRank: learning concept ranking for ontology search. Semant. Web **7**(4), 447–461 (2016). https://doi.org/10.3233/SW-150185
9. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 621–630. ACM (2009). https://doi.org/10.1145/1645953.1646033
10. Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., Corcho, O.: Ontological representation of smart city data: from devices to cities. Appl. Sci. **9**(1), 32 (2019). https://doi.org/10.3390/app9010032
11. Fernández-López, M., Suárez-Figueroa, M.C., Gómez-Pérez, A.: Ontology development by reuse. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology Engineering in a Networked World, pp. 147–170. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24794-1_7
12. Gyrard, A., Bonnet, C., Boudaoud, K., Serrano, M.: Lov4iot: a second life for ontology-based domain knowledge to build semantic web of things applications. In: IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 254–261. IEEE (2016). https://doi.org/10.1109/FiCloud.2016.44
13. Gyrard, A., Zimmermann, A., Sheth, A.: Building IoT-based applications for smart cities: how can ontology catalogs help? IEEE Internet Things J. **5**(5), 3978–3990 (2018). https://doi.org/10.1109/JIOT.2018.2854278
14. Katsumi, M., Grüninger, M.: Choosing ontologies for reuse. Appl. Ontol. **12**(3–4), 195–221 (2017). https://doi.org/10.3233/AO-160171
15. Kolbe, N., Kubler, S., Robert, J., Le Traon, Y., Zaslavsky, A.: Linked vocabulary recommendation tools for internet of things: a survey. ACM Comput. Surv. (CSUR) **51**(6), 127 (2019). https://doi.org/10.1145/3284316
16. Kolchin, M., et al.: Ontologies for web of things: a pragmatic review. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2015. CCIS, vol. 518, pp. 102–116. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24543-0_8
17. Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retrieval **3**(3), 225–331 (2009). https://doi.org/10.1007/978-3-642-14267-3
18. Martínez-Romero, M., Jonquet, C., O'Connor, M.J., Graybeal, J., Pazos, A., Musen, M.A.: NCBO ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. J. Biomed. Semant. **8**(1), 21 (2017). https://doi.org/10.1186/s13326-017-0128-y
19. McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action: Covers Apache Lucene 3.0. Manning Publications Co., Shelter Island (2010). ISBN 1933988177

20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
21. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748
22. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report 1999-66, Stanford InfoLab (1999)
23. Poveda Villalón, M., García Castro, R., Gómez-Pérez, A.: Building an ontology catalogue for smart cities, pp. 829–839. CRC Press (2014)
24. Robertson, S.E.: Overview of the Okapi projects. J. Doc. **53**(1), 3–7 (1997). https://doi.org/10.1108/EUM0000000007186
25. Sabou, M., Lopez, V., Motta, E., Uren, V.: Ontology selection: ontology evaluation on the real semantic web. In: 4th International Workshop on Evaluation of Ontologies for the Web (2006)
26. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0
27. Schaible, J., Gottron, T., Scherp, A.: Survey on common strategies of vocabulary reuse in linked open data modeling. In: Presutti, V., et al. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 457–472. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07443-6_31
28. Schaible, J., Gottron, T., Scherp, A.: *TermPicker*: enabling the reuse of vocabulary terms by exploiting data from the linked open data cloud. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 101–117. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_7
29. Simperl, E.: Reusing ontologies on the semantic web: a feasibility study. Data Knowl. Eng. **68**(10), 905–925 (2009). https://doi.org/10.1016/j.datak.2009.02.002
30. Stadtmüller, S., Harth, A., Grobelnik, M.: Accessing information about linked data vocabularies with vocab.cc. In: Li, J., Qi, G., Zhao, D., Nejdl, W., Zheng, H.T. (eds.) Semantic Web and Web Science. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6880-6_34
31. Stavrakantonakis, I., Fensel, A., Fensel, D.: Linked open vocabulary ranking and terms discovery. In: Proceedings of the 12th International Conference on Semantic Systems, pp. 1–8. ACM (2016). https://doi.org/10.1145/2993318.2993338
32. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semant. Web **8**(3), 437–452 (2017). https://doi.org/10.3233/SW-160213
33. Wu, G., Li, J., Feng, L., Wang, K.: Identifying potentially important concepts and relations in an ontology. In: Sheth, A., et al. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 33–49. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88564-1_3
34. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semant. Web **7**(1), 63–93 (2016). https://doi.org/10.3233/SW-150175