TermPicker: Enabling the Reuse of Vocabulary Terms by Exploiting Data from the Linked Open Data Cloud

An Extended Technical Report

Johann Schaible
GESIS - Leibniz Institute for the Social Sciences
Germany
johann.schaible@gesis.org

Thomas Gottron
WeST - Institute for Web Science and Technologies
Germany
gottron@gmail.com

Ansgar Scherp
ZBW – Leibniz Information Centre for Economics,
Germany
Knowledge Discovery, Kiel University, Germany
a.scherp@zbw.eu
asc@informatik.uni-kiel.de

ABSTRACT

Deciding which vocabulary terms to use when modeling data as Linked Open Data (LOD) is far from trivial. Choosing too general vocabulary terms, or terms from vocabularies that are not used by other LOD datasets, is likely to lead to a data representation, which will be harder to understand by humans and to be consumed by Linked data applications. In this technical report, we propose *TermPicker*: a novel approach for vocabulary reuse by recommending RDF types and properties based on exploiting the information on how other data providers on the LOD cloud use RDF types and properties to describe their data. To this end, we introduce the notion of so-called *schema-level patterns* (SLPs). They capture how sets of RDF types are connected via sets of properties within some data collection, e.g., within a dataset on the LOD cloud. TermPicker uses such SLPs and generates a ranked list of vocabulary terms for reuse. The lists of recommended terms are ordered by a ranking model which is computed using the machine learning approach Learning To Rank (L2R). TermPicker is evaluated based on the recommendation quality that is measured using the Mean Average Precision (MAP) and the Mean Reciprocal Rank at the first five positions (MRR@5). Our results illustrate an improvement of the recommendation quality by $29 - 36\%$ when using SLPs compared to the beforehand investigated baselines of recommending solely popular vocabulary terms or terms from the same vocabulary. The overall best results are achieved using SLPs in conjunction with the Learning To Rank algorithm *Random Forests*.

1. INTRODUCTION

When modeling Linked Open Data (LOD), engineers employ Resource Description Framework (RDF) vocabularies to represent their data as LOD. An RDF vocabulary is a collection of (unique) vocabulary terms, i.e., RDF types (also referred to as "classes") and properties, that represent a model about a certain domain. It is considered best practice to choose RDF types and properties from existing vocabularies, i.e., reuse vocabulary terms, before defining proprietary terms to create a LOD model. This reduces heterogeneity in the data representation by generating some ontological agreement with other data providers [18]. However, finding vocabulary terms that are *appropriate* for reuse is far from trivial. Prominent services, such as the, Linked Open Vocabularies catalog (LOV) [34], vocab.cc [31], and others, can be used to find specific RDF types and properties based on string search. LOV also provides an opportunity to use SPARQL for exploiting T-Box information on vocabularies and their terms, i.e., Linked Data engineers can search for equivalent RDF types or properties via owl:equivalentClass or owl:equivalentProperty, for sub-classes and sub-properties via rdfs:subClass or rdfs:subProperty, or for other relations between vocabulary terms which are defined within the vocabularies. However, these services do not exploit any A-Box information, unless a vocabulary is pointing to datasets that use the vocabulary. Services like LODstats [3] go a step further and use A-Box information to provide detailed statistics on the usage of vocabularies and vocabulary terms. However, none of these services provide information on how data providers on the LOD cloud combine the RDF types and properties from the different vocabularies to model their entire dataset.

Thus, selecting appropriate vocabulary terms for reuse can still be time-consuming, if one intends to reuse terms that other LOD providers use for publishing similar data. One has to identify such LOD providers among the amount of different datasets on the LOD cloud and examine their data on instance-level, i.e., browse through resources and examine of which RDF type they are and which outgoing properties they have.

In this paper, we introduce *TermPicker*:[1] a novel vocabulary term recommendation approach enabling the reuse of vocabulary terms by exploiting already published datasets on the LOD cloud. It provides Linked Data engineers a possibility to choose RDF types and properties used by other LOD providers in a manner that is similar to the engineers' needs.

To leverage the information how other LOD providers modeled their data, one needs to induce some structural information about which vocabulary terms were used to model entities and their relationships. In this paper, the structural information is captured by so-called *schema-level patterns* (SLPs). A schema-level pattern is a tuple describing the connection between two sets of RDF types via a set of properties. For example, the SLP

$$(\{\text{swrc:Publication}\}, \{\text{dc:creator}\}, \{\text{foaf:Person}\})$$

[1]This is a preliminary URL for the review process: `http://bit.ly/termpicker-eval`, last access 9/14/15

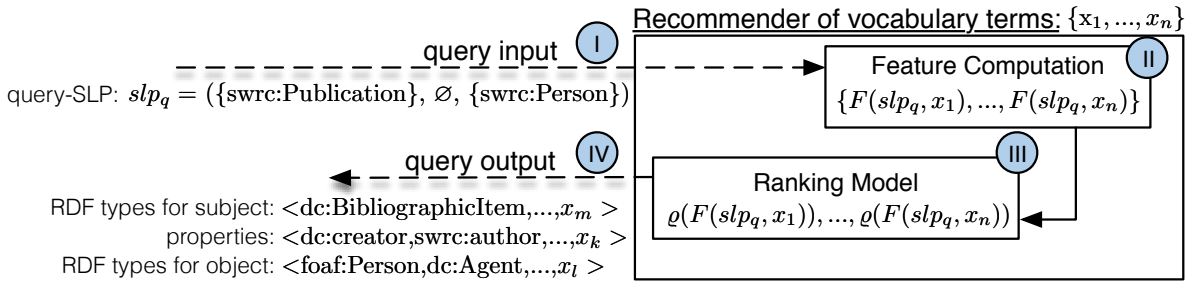

Figure 1: *Example.* A Linked Data engineer models data as LOD illustrating publications and a persons, who are the corresponding creator of a publication. She decided to reuse the SWRC vocabulary and has already chosen to use swrc:Publication and swrc:Person. TermPicker uses this information and provides her with RDF vocabulary term recommendation from other vocabularies, such as FOAF, which were used by other LOD providers along with the chosen vocabulary terms. In detail, TermPicker's input is the query-SLP $slp_q = (\{\text{swrc:Publication}\}, \varnothing, \{\text{swrc:Person}\})$ (step (I)). In step (II), the query-SLP is extended by a recommendation candidate x_i from the set $\{x_1, ..., x_n\}$ of all terms published on the LOD cloud, and five features are calculated for each extended query-SLP. The resulting feature values $F(slp_q, x_i)$ are used by the ranking model in step (III) to order all vocabulary term recommendations from most to least appropriate, before providing the ranked lists as output in Step (IV).

specifies that within one LOD collection (e.g. a dataset on the LOD cloud) resources of RDF type swrc:Publication are connected to other resources of RDF type foaf:Person via the property dc:creator. The input for TermPicker is such an SLP that is specified by the user, i.e. the query-SLP slp_q. TermPicker aims at extending the query-SLP by recommending additional vocabulary terms, which are used in other SLPs, which are calculated from existing datasets on the LOD cloud, and that are similar to slp_q.

The ranking of the recommendation candidates, i.e., vocabulary terms extracted from vocabularies published on the LOD cloud, is computed based on five features. Three of the five features represent the *popularity* of the recommendation candidate, i.e., how many data providers on the LOD cloud use the candidate, how many providers use the candidate's vocabulary, and what is the total number of occurrences of the candidate on the LOD cloud. The fourth feature specifies if the recommendation candidate is from a vocabulary that is already used in the query-SLP slp_q. Finally, the fifth feature is the so-called "SLP-feature". It calculates the number of SLPs computed from datasets on the LOD cloud, which contain all terms from slp_q as well as the recommendation candidate. In other words, the SLP-Feature investigates whether other data providers on the LOD cloud have used the recommended term in a similar SLP to slp_q, i.e., in a similar manner. The output is a set of three lists of vocabulary terms containing RDF types for resources and properties connecting these resources. These lists are ordered by a ranking model, which is induced from some training data using the machine learning approach *Learning To Rank* (L2R). Learning To Rank is a family of supervised learning algorithms to establish a ranking over a set of items, in our case vocabulary terms, by observing a general coherence between the utilized features and the relevance of an item.

Figure 1 illustrates TermPicker's general workflow and its components, such as the computation of the features and the ranking model ϱ. Let us assume, a Linked Data engineer wants to model some data as LOD illustrating publications and each publication's creator. She decided to reuse the SWRC[2] vocabulary and has already chosen to use swrc:Publication and swrc:Person. In a first step, TermPicker receives the input in form of the query-SLP $slp_q = (\{\text{swrc:Publication}\}, \varnothing, \{\text{swrc:Person}\})$ ($\varnothing$ denotes an empty set). TermPicker uses this information and provides the engineer with RDF vocabulary term recommendation from other vocabularies, such as FOAF[3], which were used by other LOD providers along with the chosen vocabulary terms. To this end, slp_q is extended with a recommendation candidate from a set of all vocabulary terms $\{x_1, ..., x_n\}$ that are published on the LOD cloud, and the five features introduced beforehand are computed for the extended query-SLP. The ranking model in the third step establishes three ranked lists of vocabulary terms that represent TermPicker's output. One list contains RDF type recommendations for the resources in subject position, another one contains the RDF type recommendations for resources in object position, and the third one comprises recommendations of properties to connect these resources. As these recommendations contain RDF types and properties from other vocabularies, the engineer is helped in finding also equivalent terms, which might better suit the engineer's need, e.g., using foaf:Person instead of swrc:Person.

We conduct a 10-fold leave-one-out evaluation to measure TermPicker's recommendation quality in different situations, in which one needs to select a vocabulary term for reuse. The recommendation quality is assessed using the Mean Average Precision (MAP) and the Mean Reciprocal Rank at the first five positions (MRR@5). As *gold standard*, we do not rely on human judgment, but rather use an automated held-out approach, i.e., before providing TermPicker with a query-SLP, we randomly extract several terms from this SLP, and solely the extracted terms are considered relevant; each other recommended term is considered irrelevant. We perform such an evaluation using data from the Billion Triple Challenge 2014 [20] as well as from the DyLDO seed-list [19] dataset. The query-SLPs for training and testing the ranking model are computed from ten different pay-level domains (PLDs), which have a relatively high ratio between reused vocabulary terms and all terms describing the data. The triples and the calculated SLPs from the remaining PLDs represent the datasets that are already published on the LOD cloud. The calculated SLPs from nine PLDs are used to train the ranking model and the calculated SLPs from one PLD are used to validate

[2] `http://swrc.ontoware.org/ontology`, last access 12/01/15

[3] `http://xmlns.com/foaf/0.1/`, last access 12/01/15

the ranking model. As the SLPs are computed from real-world data, they vary by different vocabulary terms and by the number of contained vocabulary terms. To evaluate different ranking models, we use the L2R algorithms contained in the RankLib[4] library, which provides an entire framework to train and evaluate diverse ranking models. Summarizing, the main contributions of this paper are:

(i) Evaluation of the diverse Learning To Rank algorithms contained in the RankLib library that are used to calculate a ranking model for TermPicker's recommendations.

(ii) Evaluation of the SLP-feature's impact on the recommendation quality by comparing its recommendations to the baselines of recommending solely popular vocabulary terms and recommending terms from an already used vocabulary [29, 24].

(iii) Evaluation of the different recommendations regarding whether to choose an RDF type for resources in subject position of a triple, an RDF type describing resources in object position, or to pick a property, as this reflects different real-world LOD modeling scenarios [27].

The paper is structured as follows: Section 2 describes the notion of schema-level patterns in detail and depicts how they are computed from RDF triples. Section 3 illustrates the general workflow of the proposed recommendation approach including a detailed description of the five features and a brief introduction to L2R. The evaluation of the proposed approach is described in Section 4, whereas the results of the evaluation are illustrated in Section 5. TermPicker and the evaluation results are discussed in Section 6. The related work is discussed in Section 7, in which we also differentiate TermPicker's approach to existing tools and services, before we conclude the paper.

2. SCHEMA-LEVEL PATTERNS (SLPS)

When reusing vocabularies with the goal to preferably result in some ontological agreement in data representation, one must investigate how other Linked Data providers modeled their data. Investigating solely the specification or documentation of vocabularies does not provide such information. To know which properties are used to connect resources of specific RDF types, existing datasets published on the LOD cloud must be investigated on instance level, i.e., one must browse through the data. This can be very time consuming, specifically as the number of datasets on the LOD cloud is rising.

To alleviate the situation, we introduce the notion of schema-level patterns (SLPs). They illustrate how the resources from a dataset on the LOD cloud are connected. For example, the schema-level pattern

$$slp = (\{\text{foaf:Person}, \text{dbo:ChessPlayer}\}, \{\text{foaf:knows}\}, \{\text{foaf:Person}, \text{dbo:Coach}\}) \quad (1)$$

illustrates that resources of types foaf:Person and dbo:ChessPlayer are connected to resources of types foaf:Person and dbo:Coach via the property foaf:knows. Such SLPs can be calculated from existing data sets on the LOD cloud, i.e., the SLPs are calculated based on an RDF triple representation, such as N3[5], Turtle[6], or others. The SLP in equation (1) is calculated from the fictive RDF triples in Listing 1.

[4] http://sourceforge.net/p/lemur/wiki/RankLib/, last access 9/14/15

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix dbo: <http://dbpedia.org/ontology/>

<http://ex1.org/sports_001>
   rdf:type foaf:Person;
   rdf:type dbo:ChessPlayer;
   foaf:knows <http://ex1.org/employee_002>.

<http://ex1.org/sports_002>
   rdf:type foaf:Person;
   rdf:type dbo:Coach.
```

Listing 1: Fictive RDF triples in Turtle syntax. The RDF triples specify that a resource of types Person and ChessPlayer knows a resource of types Person and Coach

Table 1: Tabular overview of the variables that are used and explained in Section 2.1 and Section 2.2

Variable	Definition
$\mathbb{V}$	Set of all vocabularies on the LOD cloud
$\mathbb{T}$	Set of all RDF types from all vocabularies in $\mathbb{V}$
$\mathbb{P}$	Set of all properties all vocabularies in $\mathbb{V}$
slp	A schema-level pattern with $slp = (sts, ps, ots)$
sts	Subject type set with $sts \in \mathcal{P}(\mathbb{T})$: RDF types describing a resource in subject position of a triple
ots	Object type set with $ots \in \mathcal{P}(\mathbb{T})$: RDF types describing a resource in object position of a triple
ps	Property set with $ps \in \mathcal{P}(\mathbb{P})$: properties interlinking resources of types in sts and ots
$\mathbb{DS}$	The set of datasets that are published on the LOD cloud
G	A graph representing a dataset such that $G \in \mathbb{DS}$
(s, p, o, c)	An RDF quadruple consisting of a subject, property, object, and a context URI where G can be found

SLPs provide an easy to use possibility for investigating how other data providers on the LOD cloud have modeled their data without having to look into the data itself. Thus, choosing vocabulary terms that are recommended based on SLPs will eventually result in an ontological agreement in data representation.

In the following we define schema-level patterns formally (cf. Section 2.1) and describe how they can be computed from existing LOD sources in Section 2.2.

2.1 Formal Definition of Schema-Level Patterns

For a better overview, the most important variables used to define SLPs are enlisted in Table 1.

Let $\mathbb{V} = \{V_1, V_2, ..., V_n\}$ be the set of all vocabularies used by datasets on the LOD cloud. Each vocabulary $V \in \mathbb{V}$ consists of vocabulary terms that are either an instance of rdfs:Class or rdfs:Property, such that $V = P_V \cup T_V$, where P_V is the set of

[5] http://www.w3.org/TeamSubmission/n3/, last access 12/01/15

[6] http://www.w3.org/TR/turtle/, last access 12/01/15

all properties p and T_V is the set of all RDF types t in vocabulary V. Accordingly, $\mathbb{T} = \bigcup_{V \in \mathbb{V}} T_V$ is the set of all RDF types and $\mathbb{P} = \bigcup_{V \in \mathbb{V}} P_V$ the set of all properties on the LOD cloud. The formal definition of an SLP is

$$slp \in \mathcal{P}(\mathbb{T}) \times \mathcal{P}(\mathbb{P}) \times \mathcal{P}(\mathbb{T}) \tag{2}$$

where $\mathcal{P}(\mathbb{T})$ is the power set of all RDF types and $\mathcal{P}(\mathbb{P})$ the power set of all properties on the LOD cloud. Based on this, an SLP is a tuple

$$slp = (sts, ps, ots) \tag{3}$$

where $sts \in \mathcal{P}(\mathbb{T})$ is the set of RDF types describing resources in subject position of a triple, $ots \in \mathcal{P}(\mathbb{T})$ the set of RDF types describing resources in object position of a triple, and $ps \in \mathcal{P}(\mathbb{P})$ the set of properties interlinking the resources of types in sts and ots.

To operate with SLPs, we define the two operators $\oplus$ and $\ominus$. The commutative $\oplus$ operator combines two SLPs:

$$slp_i \oplus slp_j := (sts_i \cup sts_j, ps_i \cup ps_j, ots_i \cup ots_j) \tag{4}$$

It can also be used for extending an SLP with a further vocabulary term by adding it either to the sets sts, ps, or ots. In detail, the operator $\oplus_{sts}$ adds an RDF type to the set sts, operator $\oplus_{ots}$ adds a RDF type to ots and the operator $\oplus_{ps}$ adds a property to the set of properties ps. This is specifically useful for examining whether a query-SLP is used in combination with a recommendation candidate by other data providers on the LOD cloud. The operation to remove terms from an SLP via the $\ominus$ is defined accordingly. The operator $\ominus_{sts}$ removes an RDF type from the set sts, operator $\ominus_{ots}$ a RDF type from ots and the operator $\ominus_{ps}$ removes a property from the set of properties ps. An example for first removing a property from an SLP and subsequently extending the SLP with an RDF type for resources in object position would be as follows:

$$\begin{aligned} slp =& (\{\text{foaf:Person}\}, \{\text{dc:date}\}, \varnothing) \ominus_{ps} \text{dc:date} \\ =& (\{\text{foaf:Person}\}, \varnothing, \varnothing) \\ slp =& (\{\text{foaf:Person}\}, \varnothing, \varnothing) \oplus_{ots} \text{foaf:Image} \\ =& (\{\text{foaf:Person}\}, \varnothing, \{\text{foaf:Image}\}) \end{aligned}$$

The relationship "$\leq$" between two schema-level patterns slp_i and slp_j illustrates that one SLP can be a *subset* of another SLP. It is defined as

$$\begin{aligned} slp_i \leq slp_j, \quad \text{iff} \quad & (sts_i \subseteq sts_j) \wedge (ps_i \subseteq ps_j) \\ & \wedge (ots_i \subseteq ots_j) \end{aligned} \tag{5}$$

and illustrates that slp_j contains more or at least as many vocabulary terms as slp_i. The strict relation $slp_i < slp_j$ defines that at least one set sts_i, ps_i, or ots_i is a proper subset of sts_j, ps_j, or ots_j, respectively. Such a relation is useful for comparing two SLPs, especially to inspect whether a query-SLP in conjunction with a recommendation candidate is a subset of other SLPs calculated from datasets on the LOD cloud.

2.2 Computing SLPs from Linked Open Data

Let $\mathbb{DS} = \{G_1, G_2, ..., G_m\}$ be the set of all data sources on the LOD cloud. Hereby, G denotes the graph of the data source and can be considered as a set of quadruples with

$$\begin{aligned} G = \{(s, p, o, c) \mid & s \in \textit{URI} \cup \textit{BN}, p, c \in \textit{URI}, \\ & o \in \textit{URI} \cup \textit{BN} \cup \textit{LIT}\} \end{aligned} \tag{6}$$

where *URI* is a set of URI's, *BN* a set of blank nodes, and *LIT* a set of literals. A triple consists of s, p, and o with s being the subject, p being the property, and o being the object of a triple. The context URI c specifies where graph G can be found. Function $\lambda(G) = \{slp_1, slp_2, ..., slp_k\}$ defines the set of SLPs that are computed from the according graph G. The specification of $\lambda : \mathbb{DS} \rightarrow \mathbb{SLP}$ is

$$\begin{aligned} \lambda(G) = \{(sts, ps, ots) \mid \exists\, s, o : \\ (\forall\, t_s \in \textit{sts} : (s, \text{rdf:type}, t_s) \in G) \wedge \\ (\forall\, p \in \textit{ps} : (s, p, o) \in G) \wedge \\ (\forall\, t_o \in \textit{ots} : (o, \text{rdf:type}, t_o) \in G)\} \end{aligned} \tag{7}$$

Hereby, $\mathbb{SLP}$ is the joint set of schema-level patterns that are computed from each graph $G \in \mathbb{DS}$

$$\mathbb{SLP} = \bigcup_{G \in \mathbb{DS}} (\lambda(G)) \tag{8}$$

An example for calculating an SLP from a graph G is provided in Equation (1), which illustrates a computed SLP from the data listed in Listing 1.

3. PICKING VOCABULARY TERMS USING SLPS

Besides illustrating how resources of specific RDF types are connected to each other, schema-level patters can be used to recommend vocabulary terms for reuse. TermPicker's input is an SLP, i.e., the query-SLP slp_q. It is extended with a vocabulary term x from the set of terms from all data sources on the LOD cloud ($x \in (\mathbb{T} \cup \mathbb{P})$). These vocabulary terms are considered to be *recommendation candidates*. Subsequently, TermPicker compares the extended query-SLP to all SLPs in $\mathbb{SLP}$. Each SLP slp_i with

$$\begin{aligned} slp_i \in \mathbb{SLP} \quad \text{and} \quad & (slp_q \oplus_{sts} x) \leq slp_i \vee \\ & (slp_q \oplus_{ps} x) \leq slp_i \vee \\ & (slp_q \oplus_{ots} x) \leq slp_i\,, \end{aligned} \tag{9}$$

is an SLP that uses vocabulary term x in combination with the terms in slp_q. Thus, vocabulary term x can be generally considered a *good* recommendation candidate for reuse. For providing meaningful recommendation candidates, the query-SLP must not be empty, i.e., $slp_q \neq (\varnothing, \varnothing, \varnothing)$, otherwise each term x would be considered a good recommendation. Also, for better readability of the paper, we generalize the extension of a query-SLP $slp_q = (sts_q, ps_q, ots_q)$ by a term x with

$$slp_q \oplus x := sts_q \cup x \vee ps_q \cup x \vee ots_q \cup x \tag{10}$$

that specifies: slp_q is extended with a vocabulary term x by adding term x either to the set sts_q, ps_q, or ots_q.

However, considering solely the existence of SLPs from $\mathbb{SLP}$ that use a recommendation candidate x in combination with the terms in slp_q, might not be sufficient to provide most reasonable recommendations. To rank each recommendation candidate from *most appropriate* to *least appropriate*, one should also encounter the popularity of the recommendation candidate and whether it is from a vocabulary that is already used in the query-SLP [29, 24]. Thus, one must first define a set of features representing each of these aspects of the recommendation candidates. A ranking model then puts the recommendation candidates in order by using these features. Establishing a general ranking model based on observing coherences between the features manually is a challenging task. Therefore,

Table 2: Overview of the utilized features. The features are computed for every recommendation candidate $x \in (\mathbb{T} \cup \mathbb{P})$

Feature	Definition
f_1	Number of datasets on the LOD cloud using the recommendation candidate x
f_2	Number of datasets on the LOD cloud using the vocabulary V_x of recommendation candidate x
f_3	Total number of occurrences of recommendation candidate x on the LOD cloud
f_4	Whether the recommendation candidate x is from a vocabulary that is already used in query-SLP slp_q
f_5	Number of SLPs in $\mathbb{SLP}$ that contain recommendation candidate x in conjunction with slp_q

TermPicker utilizes a Learning To Rank (L2R) algorithm that observes such coherences in an automatic way.

In the following, we describe and explain each feature that is used to categorize a recommendation candidate as well as the features' computation in Section 3.1. The machine learning approach Learning To Rank and how it is used to generate a ranking model for recommending vocabulary terms is briefly illustrated in Section 3.2.

3.1 Feature Computation

The set of features that categorize each recommendation candidate x is enlisted in Table 2. This set of features was derived from [29], which illustrated that the most common strategies and influencing factors to choose a vocabulary terms for reuse is its *popularity* and whether or not it is from a vocabulary that is already used. Features f_1 to f_3 represent the popularity of a vocabulary terms whereas feature f_4 specifies whether the recommended term is from a vocabulary that is is already used in the query-SLP. Additionally, we introduce feature f_5 that calculates how many SLPs $slp_i \in \mathbb{SLP}$ exist with $slp_q \oplus x \leq slp_i$. Each of these features represent some factor that an engineer might consider important in her vocabulary term choice. However, none of the features encode the relevance of a recommendation candidate directly. In Sections 3.1.1 to 3.1.3, these five features are described in detail including the formalizations for their computation.

3.1.1 Popularity (Features f_1 to f_3)

Feature f_1 comprises the number of datasets $G \in \mathbb{DS}$ on the LOD cloud using a recommendation candidate x. It is calculated by examining whether the term x is contained in an RDF triple/quadruple of a graph G.

$$f_1(x) = |\{G \mid (\exists\,(s,p,o,c) \in G : p = x) \vee (\exists\,(s, \text{rdf:type}, o, c) \in G : o = x)\}| \tag{11}$$

Feature f_2 depicts the number of datasets on the LOD cloud using the vocabulary V_x of the recommendation candidate x. It is calculated similar to feature f_1, but it examines whether the vocabulary of term x is used in a triple of graph G.

$$f_2(V_x) = |\{G \mid (\exists\,(s,p,o,c) \in G : p \in V_x \vee (o \in V_x \wedge p = \text{rdf:type}))\}| \tag{12}$$

The total number of occurrences of the recommendation candidate x on the LOD cloud is calculated by feature f_3. In contrast to the features f_1 and f_2, feature f_3 is calculated by counting each triple/quadruple, in which the vocabulary term x is contained.

$$f_3(x) = \sum_{G \in \mathbb{DS}} |\{(s,p,o,c) \in G \mid (p = x) \vee (o = x \wedge p = \text{rdf:type})\}| \tag{13}$$

Combined, these three features define the popularity of a vocabulary term on a very fine-grained level. Whereas the total number of occurrences of a recommendation candidate x depicts its overall usage, the number of data sources using x and its vocabulary specifies whether its usage is spread across many datasets on the LOD cloud or concentrates on only a few ones. We do not normalize any of the feature values, but rather use the absolute values, as this ensures that valuable information would not be lost, i.e., normalizing the feature values in our L2R based evaluation setup could lead to false recommendations.

The benefit of reusing popular vocabulary terms is supposed to enable an easier consumption of the data, as many Linked Data consumption tools provide tailored support for popular vocabularies [18]. This is also backed up by the recommendations of the W3C when modeling LOD.[7] In addition, it makes the data more understandable for humans. TermPicker makes use of these features, as they are also acknowledged by Linked Data practitioners in a survey on their strategies and influencing factors to reuse a vocabulary term or not [29].

3.1.2 Same Vocabulary (Feature f_4)

Feature f_4 indicates whether the vocabulary of a recommendation candidate x is already contained in the query-SLP, i. e., $slp_q = (sts_q, ps_q, ots_q)$. The calculation returns a binary value, where 1 denotes that the vocabulary of term x is already used in slp_q, and 0 if it is not contained in slp_q.

$$f_4(slp_q, x) = \begin{cases} 1 & \text{if } \exists\, V : x \in V \wedge \\ & (sts_q \cup ps_q \cup ots_q) \cap V \neq \emptyset \\ 0 & \text{else} \end{cases} \tag{14}$$

Reusing terms from the same vocabulary is considered as an important strategy not only in the survey on vocabulary reuse strategies described in [29], but specifically in certain domains such as the statistics domain. There, it is accustomed to reuse primarily vocabulary terms from SKOS[8] or XKOS[9] [24]. In other words, one might want to search for vocabularies covering the domain of interest and subsequently adapt RDF types and properties from those vocabularies for particular needs. The reason for that: it seems quite likely that one specific domain vocabulary, such as SKOS, contains many RDF types or properties that can be reused for describing data from that specific domain. Furthermore, reusing terms from the same vocabulary reduces the overload of too many different vocabularies and makes the data easier to understand for humans that are familiar with the domain specific vocabulary [29].

3.1.3 The SLP-Feature (Feature f_5)

[7] `http://www.w3.org/TR/ld-bp/#VOCABULARIES`, last access 12/12/15

[8] `http://www.w3.org/2004/02/skos/`, last access 09/06/15

[9] `http://rdf-vocabulary.ddialliance.org/xkos.html`, access 09/06/15

The SLP-feature is calculated based on a query-SLP slp_q that is extended with a recommendation candidate x. The extended query-SLP $slp_q \oplus x$ is compared to existing SLPs $slp_i \in \mathbb{SLP}$, in order to find SLPs with $(slp_q \oplus x) \leq slp_i$. The number of SLPs $slp_i \in \mathbb{SLP}$ with $(slp_q \oplus x) \leq slp_i$ represents how often other datasets on the LOD cloud use vocabulary term x in conjunction with the terms in slp_q.

$$f_5((slp_q \oplus x), \mathbb{SLP}) = |\{slp_i \mid slp_q \oplus x \quad \leq slp_i\}| \qquad (15)$$

Using recommendations based on this feature is likely to result in reducing heterogeneity in the data representation by relying on ontological agreement. The more SLPs in $\mathbb{SLP}$ use the recommendation candidate x in such a *similar* way, the more appropriate does it seem to reuse this term in order to eventually result in some ontological agreement.

3.2 Learning to Rank

Combined, features f_1 to f_5 describe each recommendation candidate x in a unique way. However, it remains unclear how these features can be used to provide a ranked list of recommendations. The feature values for each recommendation candidate might vary a lot, as in the following fictive example:

- $(slp_q \oplus x_1) =$ f_1: 7, f_2: 9, f_3: 20, f_4: 1, f_5: 4
- $(slp_q \oplus x_2) =$ f_1: 3, f_2: 3, f_3: 5, f_4: 0, f_5: 6
- $(slp_q \oplus x_3) =$ f_1: 10, f_2: 30, f_3: 80, f_4: 0, f_5: 2
- $(slp_q \oplus x_4) =$ f_1: 4, f_2: 20, f_3: 29, f_4: 1, f_5: 0

Immediately, the question arises which of these four recommendation candidates can be considered the most appropriate term for reuse. To rank these terms from most to least appropriate, one must observe a general coherence between the features and the relevance of each recommendation candidate. However, observing such a coherence manually can be quite difficult. Rather, it must be observed in an automatic way to learn the feature's impact on the quality of the recommendations.

In order to address this challenge, TermPicker makes use of the machine learning approach "Learning To Rank" (L2R). Learning to rank refers to a class of supervised machine learning techniques for inducing a ranking model [23, 17]. In detail, a ranking model ϱ allows for determining relevant and irrelevant items for a given information need. In our case, an information need corresponds to the query-SLP slp_q. The relevant and irrelevant items correspond to the recommendation candidates $x \in \mathbb{T} \cup \mathbb{P}$. The ranking model ϱ is derived from some training data by observing the mentioned general coherence between the feature values and the relevance of a recommendation candidate. Ideally, the derived ranking model lists all relevant vocabulary terms high and before less relevant or irrelevant vocabulary terms.

Formally, the ranking model ($\varrho(F(slp_q, x))$) calculates a ranking score for the recommendation candidate x, where $F(slp_q, x)$ denotes the calculation of features f_1 to f_5 for the extended query-SLP slp_q by the recommendation candidate x. This way, each recommendation candidate $x \in \mathbb{T} \cup \mathbb{P}$ can be ranked based on the ranking score in descending order. To establish such a ranking model, one needs training data to derive a general coherence between the feature values and the relevance of a recommended term. In our case, the training data is a set of query-SLPs with existing relevance information on each recommendation candidate. It contains SLPs such as

$$slp_q = (\{\text{swrc:Publication}\}, \varnothing, \{\text{foaf:Agent}\})$$

with the relevance information that e.g. for recommending properties solely the terms dc:creator and swrc:author are considered as relevant. Using this information, an L2R algorithm iterates through the training data to detect the beforehand mentioned coherence between the feature values and the relevance, such that the relevant terms get ranked as high as possible. This way, the learned ranking model can be used in new and previously unknown situations with new and unknown query-SLPs. For example, a query-SLP that was not part of the training set using terms from the Creative Commons[10] ontology and from an ontology for managing presentations at W3C[11]

$$slp_q = (\{\text{cc:Work}\}, \{\text{w3:presenter}\}, \varnothing)$$

can get recommendations, such as the RDF types foaf:Person and/or dc:Agent to reuse for resources in object position.

L2R algorithms are categorized in three different groups according to their method for learning a ranking model [23]: (A) *point-wise* L2R algorithms, (B) *pair-wise* L2R algorithms, and (C) *list-wise* L2R algorithms. A point-wise approach ranks vocabulary terms directly by allocating a ranking score to each recommendation candidate individually. Pair-wise methods rank vocabulary terms solely in a given pair of two recommendation candidates. This way, a term is considered a better recommendation compared to the terms in a lower ranking position. List-wise approaches rank recommendation candidates by optimizing the quality measure of the result list, such as the Mean Average Precision (MAP). They examine which coherence between the features provides the highest measure, e.g., the highest MAP value, and use the derived ranking model assuming the quality measure is as high in new situations.

In particular, the pair-wise and list-wise approaches have demonstrated good performance in generic ranking scenarios [6]. However, it is of interest for our use-case to determine which of the approaches, i.e., point-wise, pair-wise, or list-wise, perform better in our setting of recommending vocabulary terms for reuse.

4. EVALUATION

The proposed approach is evaluated using a 10-fold leave-one-out evaluation. Each fold comprises a *training set* to induce the ranking model, a *test set* to evaluate the ranking model, and a set representing datasets that are already published on the LOD cloud to calculate features f_1 to f_5. We investigate different ranking models and thus TermPicker's recommendation quality based on the aspects that depict the main contribution of this paper:

(i) Comparison of all Learning To Rank algorithms contained in the RankLib library that provides a framework for inducing and evaluating a ranking model. The three most competitive Learning To Rank algorithms are examined in detail, i.e., in our evaluation these three algorithms were *Coordinate Ascent* [26], *LambdaMART* [36], and *Random Forests* [4].

[10] `http://creativecommons.org/ns#`, last access 09/06/15

[11] `http://www.w3.org/2004/08/Presentations.owl#`, last access 09/06/15

(ii) Comparison of using the SLP-feature (f_5) to using features $f_1 - f_3$ (baseline of reusing only popular vocabulary terms) [29] and to using features $f_1 - f_4$ (baseline of reusing popular vocabularies from the same vocabulary) [24] to investigate the impact of the SLP-feature on the recommendation quality.

(iii) Comparison of recommending RDF types for resources in subject position of a triple, RDF types describing resources in object position, and recommending properties, as this reflects different real-world LOD modeling scenarios [27].

The recommendation quality is measured using the Mean Average Precision (MAP) and the Mean Reciprocal Rank at the first five positions (MRR@5).

In the following, Section 4.1 describes the evaluation design in detail. It is illustrated how the relevance of a recommendation candidate is defined, in order to enable the L2R algorithm to learn the ranking model. In Section 4.2 it is explained which data was used for the evaluation as well as how it was split to train and evaluate the ranking model. It also includes statistics on the data and the ten folds. Finally, we formalize the quality measures MAP and MRR@5 to illustrate how the recommendation quality was calculated.

4.1 Evaluation Design

TermPicker's recommendations are evaluated by simulating a search for an appropriate vocabulary term that can be reused. Thus, the training set and test set, which are used to induce and evaluate the ranking model, are disjunct sets of distinct SLPs. These SLPs are used as input for TermPicker. However, before providing TermPicker with these SLPs as input, one or more random vocabulary terms are extracted from that SLP using the $\ominus$ operator. These extracted terms determine the set of *relevant* recommendation candidates, as they are the ones that have been initially used. All other recommendation candidates that are not contained in the set of the extracted terms are considered as irrelevant recommendations. This way, for each query-SLP, the ranking model is provided (a) a set of recommendation candidates, (b) five feature values categorizing each recommendation candidate, and (c) the relevance of each recommendation candidate. The L2R algorithm uses this information and observes a general coherence between the feature values and the relevance of a recommendation [17].

For example, given an SLP slp_j from the training or test set with

$$slp_j = (\{\text{swrc:Publication}\}, \{\text{swrc:author}\}, \{\text{swrc:Person}\})$$

the property swrc:author is randomly extracted via the $\ominus_{ps}$ operator.

$$\begin{aligned} slp_q &= slp_j \ominus_{ps} \text{swrc:author} \\ &= (\{\text{swrc:Publication}\}, \varnothing, \{\text{swrc:Person}\}) \end{aligned}$$

The query-SLP slp_q is now provided as input for TermPicker. The output is a set of vocabulary terms, including a set of properties. The previously extracted property swrc:author is considered a relevant recommendation, as it was initially used in slp_j. Every other recommendation is considered irrelevant, as these terms were not used in slp_j. This makes it possible to induce and evaluate a ranking model by interpreting a ranked list of recommendations

$$< \text{dc:date}, \text{dc:title}, \textbf{swrc:author}, \ldots >$$

in the following way: the first two recommendations are irrelevant, and the first relevant recommendation is at the third rank of the result list.

Such an evaluation can be performed fully automatically reflecting many different real-life scenarios. Human assessment whether a recommendation is relevant or not is not required. This helps drastically to establish a first generalized ranking model using a lot of data. Relying on human judgment would be very time consuming and difficult, as the manual assessment would take a lot of time and one would need many different domain experts, in order to correctly judge every recommendation candidate. The real-life scenarios are represented by the many different query-SLPs. Each query-SLP represents the Termpicker's input provided by the engineer, and the previously extracted term represents what the engineer is looking for. Every recommendation candidate is assigned its feature values. Sometimes the previously extracted term is used by other LOD providers in conjunction with the query-SLP and sometimes not, which is reflected by the SLP feature value. Thus, the SLP feature is only an indicator that a recommendation candidate might be relevant, and therefore, the automatic evaluation provides every aspect in order to evaluate how much influence the SLP feature actually has on the recommendation quality.

4.2 Datasets for the Evaluation

To validate TermPicker's recommendation quality, we perform two separate evaluations. One evaluation uses the seed-list data of the Dynamic Linked Data Observatory (DyLDO) [19][12] and the other evaluation uses the Billion Triple Challenge dataset (BTC) 2014 [20][13] (crawl no. 1). We chose these two data sets, as they represents parts of the LOD cloud in different way. For once, DyLDO's seed list, i.e., the set of URIs that form the core of the data crawling, is different from the seed list of the BTC 2014 dataset. The seed list of the BTC 2014 dataset is sampled from the previous year's dataset, and the initial seed-list was gathered from various semantic search engines. DyLDO's seed list comprises the 220 most popular URIs selected from the BTC 2011[14] dataset based on a PageRank analysis combined with another 220 URIs from the CKAN/LOD[15] registry, which were not contained in the BTC 2011 dataset. This means that the DyLDO and the BTC 2014 datasets contain different data, as it was crawled from different dataset on the LOD cloud. Furthermore, DyLDO's seed list makes about 50% of the entire data contained in the DyLDO dataset, whereas the seed list of BTC 2014 makes only less than one percent, resulting in way more data than in DyLDO.

DyLDO comprises a considerable amount of about 10.8 million triples from 382 different pay-level domains. In total there are about 2.3 million distinct vocabulary terms from about 600 vocabularies. The BTC 2014 dataset contains about 1.4 billion triples, of which we use solely 34 millions, in order to keep the in-memory SLP computation maintainable. These 34 million triples are provided by 3,493 pay-level domains. Within these triples there are about 5.5 million distinct RDF types and properties from about 1,530 different vocabularies.

[12] http://swse.deri.org/dyldo/, last access 12/12/15
[13] http://km.aifb.kit.edu/projects/btc-2014/, last access 12/12/15
[14] https://km.aifb.kit.edu/projects/btc-2011/, last access 12/12/15
[15] https://datahub.io/dataset?tags=lod, last access 12/12/15

Table 3: PLDs that were selected as test and training in the evaluations. The selection was based on $C1$ (PLDs that provided the highest number of distinct vocabulary terms) and $C2$ (PLDs with the highest ratio between the reused vocabulary terms and all RDF types and properties). The left half of the table shows the selected PLDs from the DyLDO dataset, whereas the right half shows the selected PLDs from the BTC 2014 dataset

DyLDO				BTC 2014			
PLD	$(C1)$	$(C2)$	# of SLPs	PLD	$(C1)$	$(C2)$	# of SLPs
kasei.us	100	1.00	121	b4mad.net	291	1.00	393
thefigtrees.net	89	1.00	102	derby.ac.uk	137	1.00	197
bblfish.net	82	0.99	150	heppnetz.de	121	1.00	199
wikier.org	96	1.00	133	ivan-herman.net	196	1.00	303
bl.uk	102	0.46	246	jones.dk	164	1.00	155
kanzaki.com	176	0.99	294	ldodds.com	115	1.00	125
taxonconcept.org	139	0.92	424	lmco.com	128	1.00	204
fundacionctic.org	110	0.97	390	mfd-consult.dk	192	1.00	315
data.gov.uk	258	0.93	920	mit.edu	174	0.96	293
bbc.co.uk	146	1.00	522	nickshanks.com	100	0.97	164

We regard a vocabulary simply by its URI namespace which is specified by the W3C.[16] This means that a vocabulary either uses a *hash namespace* or a *slash namespace*, i.e., the vocabulary of a term is represented by the URI before the last occurrence of either a hash or a slash. Therefore, we do not distinguish between a vocabulary being of type owl:Ontology, voaf:Vocabulary, or others, and keep it as simple as possible. To differentiate which dataset is under the control of which data publisher, we make use of the the pay-level domain (PLD) calculated from the the context-URI c contained in the data. A pay-level domain (PLD) is a direct subdomain of a top-level domain, such as *.org* or *.com*, or of a second-level country domain, such as *.de* or *.uk*.[17] Examples of pay-level domains included in the BTC 2014 (about $3,500$ PLDs) and the DyLDO (about 382 PLDs) dataset are dbpedia.org or bbc.co.uk.

A fully qualified domain name, such as the context-URI itself, would over-exaggerate the diversity of the data, as it would also differentiate data from different sub-domains. Hence, by referring to a *dataset published on the LOD cloud* or *a data publisher on the LOD cloud*, we refer to a PLD that specifies which data publisher is in control of the data.

For each evaluation, the evaluation dataset is split by the pay-level domains. The data from ten different PLDs is used as training and test set, whereas the data from the remaining PLDs is used to simulate the data sets published on the LOD cloud. For each fold of the 10-fold leave-one-out evaluation, one of the ten PLDs is left out and resembles the test set, whereas the other nine PLD represent the training set. As mentioned before, both the test and training set consists of the computed SLPs from the data of the according pay-level domain(s).

The more query-SLPs are used to train and test the ranking model, and the larger the data for calculating the features values, the more representative are the generated results [6]. Thus, the ten pay-level domains for training and testing are selected based on two criteria.

($C1$) A high number of distinct vocabulary terms within a PLD

($C2$) A high ratio between the reused vocabulary terms and all RDF types and properties used within a PLD

The high number of distinct vocabulary terms indicates that resources of various RDF types are interlinked via several different properties. This way, it is very likely to calculate a high number of distinct SLPs from that data. A negative example is a dataset modeling several million instances of type foaf:Person knowing other persons, as this will generate solely one SLP. The high ratio between the reused terms and all terms used to describe the data indicates that most resources and their interlinking are described via reused and not self-defined vocabularies. This enables to calculate SLPs that most likely contain many reused terms, which is important to generate valuable recommendations. Selecting PLDs for training and test sets randomly and not based on $C1$ and $C2$ is very likely to result in poor evaluation results, as many PLDs either do not use many different vocabulary terms or they use many self-defined terms.

Table 3 provides an overview of the selected PLDs used for the evaluations based on the DyLDO (left half of the table) and BTC 2014 (right half of the table) dataset as well as the numbers considering $(C1)$ and $(C2)$. Those PLDs that provided the highest numbers in both $C1$ and $C2$ were selected as test and training sets. Furthermore, it displays the number of distinct SLPs that are calculated from the data of the selected pay-level domains.

Naturally, SLPs that are used to train the ranking model are different to the SLPs that are used to evaluate the model. The data from the remaining PLDs that is used for calculating the features contains $117,776$ (DyLDO) and $227,010$ (BTC 2014) SLPs, respectively.

4.3 Evaluation Metrics

As a user, who searching for possible RDF types and properties for reuse, is likely to browse only through the top-k vocabulary terms (where k is generally a small number such as 5 or 10), it is important to evaluate the ranking model by measures that use ordered sets of vocabulary terms. We use the Mean Average Precision (MAP) and the Mean Reciprocal Rank to the fifth position (MRR@5). Both measures illustrate the quality of the ranking model well, as they compute values using such ordered sets of vocabulary terms (in contrast to basic measures such as precision and recall).

[16] `http://www.w3.org/2001/sw/BestPractices/VM/http-examples/2006-01-18/#naming`, last access 9/25/15

[17] To calculate the pay-level domain, we make use of the Google guava library: `https://code.google.com/p/guava-libraries/`, last access 9/28/15

On one hand, MAP provides a measure of quality across recall levels [25]. It illustrates the quality of the entire result list in which the ranking position of the relevant vocabulary term is considered. The higher the MAP value, the more relevant vocabulary terms are ranked to the top positions of the result list. On the other hand, the Mean Reciprocal Rank at the first k results (MRR@k) investigates the result list only to the rank position of the first relevant vocabulary term [11]. In other words, MRR returns a metric specifying the ranking position of the first relevant term.

In the following, we use $k = 5$. We define the set of query-SLPs as $Q = \{slp_{q_1}, ..., slp_{q_n}\}$. If the set of relevant vocabulary terms for a query $slp_{q_j} \in Q$ is $\{rt_1, \ldots, rt_{m_j}\}$ and R_{jh} $(1 \leq h \leq m_j)$ is the set of ranked retrieval results from the top result until one gets to the relevant vocabulary term rt_h, then the **Mean Average Precision** and the **Mean Reciprocal Rank** of Q defined as

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{h=1}^{m_j} \text{Precision}(R_{jh}) \tag{16}$$

$$\text{MRR}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{|R_{jh}|} \tag{17}$$

5. RESULTS

The results of the evaluation are presented in Figure 2 and Figure 3. They illustrate the recommendation quality via box-plots based on the MAP and the MRR@5 respectively. The figures depict the measurements of the recommendation quality considering the aspects (i), (ii), and (iii) introduced in Section 4. The three most competitive L2R algorithms in the RankLib library are: *Coordinate Ascent*, *LambdaMART* and the *Random Forest* algorithm. The difference between these three L2R algorithms can be observed by comparing the three different rows in Figures 2 and 3. The varying recommendation quality between the different set of features can be examined by comparing the three columns of the Figures. Both reusing solely popular vocabulary terms (marked as POP) and reusing vocabulary terms from the same vocabulary (marked as SAME) resemble the baseline, as they are considered current state of the art strategies to reuse a vocabulary [29, 18]. Our proposed approach, marked as "SLP-feature-based", additionally uses the SLP-feature. Within each plot, the x-axis displays the different recommendations of a RDF type for resources in subject position (abbreviated as "sts"), of a RDF type for resources in object position (abbreviated as "ots"), or of a property (abbreviated as "ps") for both the BTC 2014 and the DyLDO dataset. Each box plot comprises the measured recommendation quality of the ten PLDs that were used as test sets in the 10-fold leave-one-out evaluation. The plot that is marked bold illustrates the configuration, i.e., which features and which L2R algorithm, achieving the overall best recommendation quality.

(i) Differences between L2R algorithms. Comparing the three most competitive L2R algorithms, one can observe that there are no obvious differences between the algorithms when using solely features $f_1 - f_3$ (baseline POP) or when using features $f_1 - f_4$ (baseline SAME). The median MAP and MRR@5 values are between 0.3 and 0.5 for each of the three algorithms. However, when making use of all features including the SLP-feature, the differences of the median values are more noticeably. While the median values using the algorithms *Coordinate Ascent* and *Random Forests* on the BTC 2014 data are between 0.7 and 0.8, the median values using *LamdaMART* vary in average at 0.6. Four other algorithms from the RankLib library, i.e., *AdaRank*[37], *RankNet*[5], *RankBoost* [16], and *ListNet* [7], did not provide such good results. The median MAP and MRR@5 values were never above 0.3, and there was no increase of the recommendation quality between using the different sets of features. Finally, the L2R algorithm *MART* [5] was able to achieve a median MAP and MRR@5 value of about 0.5, but in total, MART's successor, i.e., LambdaMART, provided very similar but slightly better results. These results can be observed when using the BTC 2014 dataset as well as the DyLDO dataset as evaluation data.

(ii) Impact of the SLP-feature. Comparing the different set of utilized features, one can observe that the differences are more visible when using the BTC 2014 dataset as evaluation data. There is a slight increase in the recommendation quality, when adding feature f_4 to the set of features, i.e., the medians for the baseline POP and the baseline SAME differ in average by about 7%. When adding the SLP-feature however, the median recommendation quality increases by about 30% compared to the baseline of reusing solely popular vocabulary terms (compared to POP). Even compared to the SAME baseline, i.e., reusing popular vocabulary terms from the same vocabulary, one can perceive an increase of the recommendation quality by 20%. Such differences between the sets of utilized features are not as visible when performing the evaluation on the DyLDO dataset. However, one can still observe that there is only a small increase of the recommendation quality ($< 7\%$) of the baseline SAME compared to the baseline POP. Using also the SLP-feature increases the median recommendation quality by about $15 - 20\%$ compared to the baselines POP and SAME.

(iii) Differences between recommendation types. Finally, using all features (including the SLP-feature) and comparing the recommendation quality between recommending RDF types for resources in subject position, RDF types in object position, or properties, only slight changes (between $5-10\%$) in the recommendation quality can perceived. Solely the L2R algorithm *LambdaMART* based on the BTC 2014 dataset has a higher median recommendation quality when suggesting RDF types for resources in object position (MAP $= .83$) compared to the medians when suggesting a property (MAP $= .63$) and when suggesting an RDF type for a resource in subject position (MAP $= .5$). The MRR@5 values are very much the same.

In addition, Table 4 and Table 5 illustrate the average MAP and MRR@5 values (including the standard deviation) for the evaluations based on the BTC 2014 and the DyLDO datasets, respectively. They underline the increase of the recommendation quality, when adding the SLP-feature to the set of features, which is used by the ranking model. For the BTC 2014 dataset, in average, using the SLP-feature provides a higher MAP and MRR@5 value than using features to reusing terms from the same vocabulary (SAME) by 29%, and comparing to the features for reusing solely popular vocabulary term (POP), it provides better recommendations by 36%. For the DyLDO data, these differences are not as distinctive, but they are still 13% compared to the baseline SAME and 23% compared to the baseline POP. Looking at Table 4, the L2R algorithm *Coordinate Ascent* seems to provide the best results with a MAP of MAP $= .76$ and an MRR@5 value of MRR@5 $= .81$. However, it does not perform as well based on the DyLDO dataset (MAP $= .43$ and MRR@5 $= .55$). Therefore, the overall best recommendation quality, which is calculated based on the values from

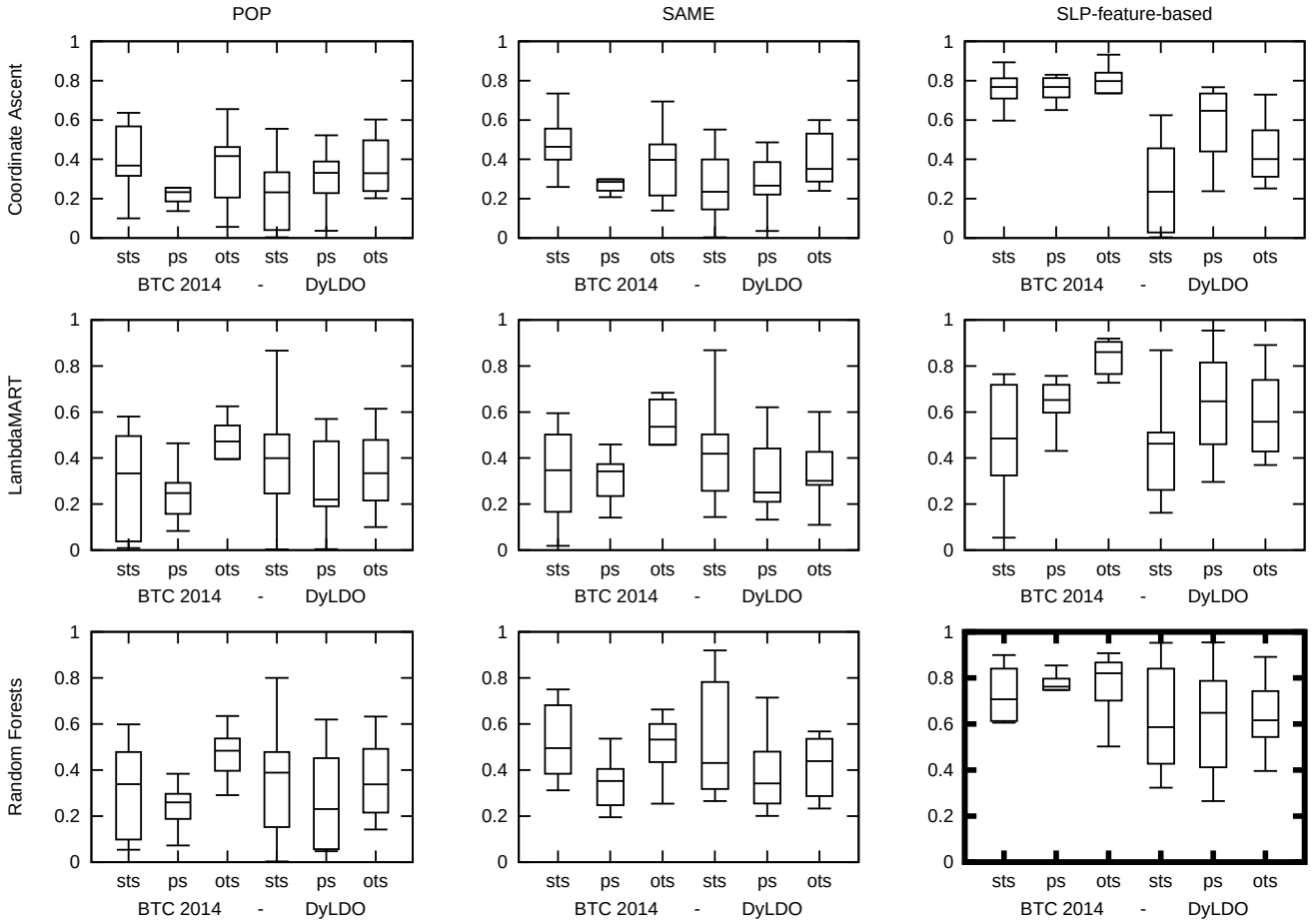

Figure 2: MAP results. On the x-axis of each plot one finds the recommendations for RDF types for resources in subject position "sts", for properties "ps", of for RDF types for resources in object position "ots". The left part of each plot represents the results of the evaluation performed on the BTC 2014 dataset and the right part of the plots depicts the results using the DyLDO dataset. The proposed SLP-Feature can be compared with the baseline reusing popular vocabularies (POP) and the baseline reusing popular vocabularies from the same vocabulary (SAME) for the three most competitive L2R algorithm from the RankLib library. The plot marked bold depicts the overall best results, which is the Random Forests algorithm using the SLP-Feature.

Table 4 and Table 5, is provided by the L2R algorithm *Random Forests* using all features, including the SLP feature (MAP $= .70$ and MRR@5 $= .73$).

6. DISCUSSION

The discussion is structured as follows: In Section 6.1, we discuss the results of the evaluation based on the three main contributions of this paper, i.e., (i) the difference between the utilized Learning To Rank algorithms, (ii) the impact of the SLP-feature on the recommendation quality, and (iii) the difference between recommending RDF types and properties. We also provide insights whether the measured differences are significant using the Friedman test (differences are significant with a p-value $p < .05$) and a Wilcoxon signed-rank test with a Bonferroni correction applied to detect pair-wise differences (the corrected p-value for (i) to (iii) is $p < (.05/3 = .017)$). In Section 6.2, we discuss the general use of a Learning to Rank algorithm for providing vocabulary term recommendations, as well as the limitations of the utilized evaluation design.

6.1 Discussion of the Results

6.1.1 Differences between the L2R algorithms (i)

From the eight L2R algorithms contained in the RankLib library, solely four algorithms were able to provide recommendations with an MAP above 50% when making use of all features. Out of the four algorithms with MAP < 0.5, two algorithms (*RankNet* and *RankBoost*) are pair-wise approaches, and the other two algorithms (*ListNet* and *AdaRank*) are list-wise approaches. The best performing algorithm, i.e., *Random Forests*, is a point-wise approach, whereas the other ones (*Coordinate Ascent*, *LambdaMART*, and *MART*) are all list-wise approaches.

Generally, list-wise and pair-wise approaches perform better than point-wise approaches [1, 6]. However, in cases where there is only a binary relevance, i.e., a recommendation candidate is either relevant or irrelevant, point-wise approaches perform better, if there is solely one relevant recommendation candidate for most queries [1, 6]. In our use-case, recommendation candidates have indeed a binary relevance. Additionally, most of the query-SLPs used to train and evaluate the ranking model contained mostly up to three

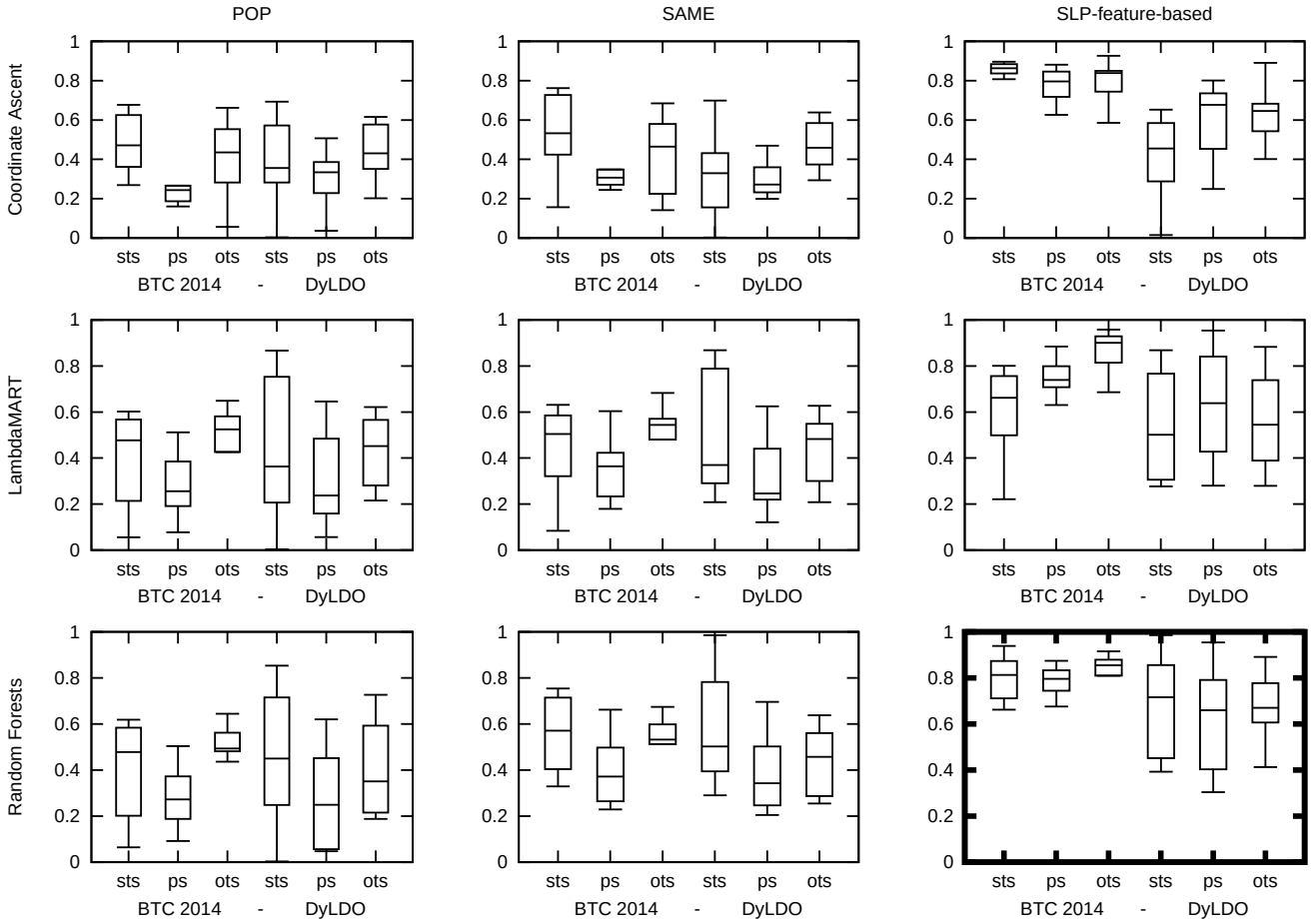

Figure 3: MRR@5 results. On the x-axis of each plot one finds the recommendations for RDF types for resources in subject position "sts", for properties "ps", of for RDF types for resources in object position "ots". The left part of each plot represents the results of the evaluation performed on the BTC 2014 dataset and the right part of the plots depicts the results using the DyLDO dataset. The proposed SLP-Feature can be compared with the baseline reusing popular vocabularies (POP) and the baseline reusing popular vocabularies from the same vocabulary (SAME) for the three most competitive L2R algorithm from the RankLib library. The plot marked bold depicts the overall best results, which is the Random Forests algorithm using the SLP-Feature.

vocabulary terms. Therefore, based on the evaluation design, only one or two vocabulary terms could be extracted, to provide relevant recommendation candidates and to provide TermPicker with a non-empty query-SLP. Thus, in our evaluation, we use a binary relevance, and for most of the queries there are solely one or two relevant recommendation candidates. Based on this, it is quite reasonable that a point-wise L2R algorithm performs best.

This is underlined by the significant differences between the recommendation quality using the algorithm *Random Forests* and the recommendation quality using the other L2R algorithms. The Friedman test, which compares the overall MAP and MRR@5 values based on both the BTC 2014 and the DyLDO data using all features, showed that these differences are statistically significant with $\mathcal{X}^2 = 14,000, p = .001$. The Wilcoxon signed-rank with Bonferroni correction applied proved that there is no significant difference between using the *Coordinate Ascent* and *LambdaMART* algorithm ($Z = -0.243$, $p = .808$ **n.s.**). However, with $Z = -2.492, p = .013$, the *Random Forests* algorithm provides significantly better recommendation than the *Coordinate Ascent* algorithm, and with $Z = -4.237, p < 0.001$ it is also significantly better than *LambdaMART*.

6.1.2 *Impact of the SLP-feature (ii)*

To discuss the impact of the features on the recommendation quality, we use the best performing L2R algorithm for each set of features across both the BTC 2014 and the DyLDO dataset, i.e., for the baseline POP that is the L2R algorithm *LambdaMART* and for the baseline SAME as well as for using the SLP-feature that is the algorithm *Random Forests*.

With MAP $\approx$ .35, the average MAP value of recommendations based on reusing solely popular vocabulary terms (baseline POP) is quite high. Specifically considering the fact, that the feature values describing the popularity of a recommendation candidate are static, meaning they do not depend on the query-SLP. However, such MAP values can be explained by the setup of our evaluation. As we use real-life data for our evaluation, the relevant recommendation candidates are vocabulary terms that actually have been used by some ontology engineer to describe the data. The best prac-

Table 4: MAP and MRR@5 values for BTC 2014. Each row depicts the average MAP and MRR@5 values and their standard deviation for the three most competitive L2R algorithms in the RankLib library and the set of features, i.e., baseline POP, baseline SAME, and using the SLP-feature. The columns depict the difference between recommending a property or RDF types for resources at subject or object positions of a triple. The overall recommendation quality of a L2R algorithm with a specific set of features is illustrated in the two most right columns

		sts		ps		ots		overall	
Model	Featues	MAP	MRR@5	MAP	MRR@5	MAP	MRR@5	MAP	MRR@5
CoordinateAscent	POP	.38 (.18)	.49 (.15)	.25 (.11)	.27 (.11)	.38 (.19)	.41 (.19)	.34 (.16)	.39 (.15)
	SAME	.48 (.16)	.55 (.19)	.31 (.10)	.33 (.09)	.39 (.18)	.43 (.18)	.39 (.15)	.44 (.15)
	SLP	.75 (.12)	.83 (.10)	.76 (.06)	.78 (.08)	.76 (.14)	.81 (.10)	.76 (.11)	.81 (.09)
LambdaMART	POP	.31 (.22)	.39 (.21)	.27 (.15)	.28 (.14)	.42 (.20)	.45 (.20)	.33 (.19)	.37 (.18)
	SAME	.34 (.21)	.44 (.17)	.33 (.13)	.34 (.14)	.49 (.19)	.49 (.17)	.39 (.18)	.42 (.16)
	SLP	.46 (.25)	.61 (.18)	.64 (.10)	.73 (.12)	.82 (.12)	.86 (.09)	.64 (.16)	.73 (.13)
RandomForests	POP	.32 (.20)	.40 (.21)	.26 (.12)	.28 (.12)	.45 (.17)	.48 (.15)	.34 (.16)	.39 (.16)
	SAME	.52 (.16)	.56 (.15)	.37 (.14)	.39 (.14)	.49 (.16)	.50 (.17)	.46 (.15)	.48 (.15)
	SLP	.72 (.11)	.80 (.10)	.75 (.10)	.77 (.10)	.78 (.12)	.83 (.08)	.75 (.11)	.8 (.09)

Table 5: MAP and MRR@5 values for DyLDO. Each row depicts the average MAP and MRR@5 values and their standard deviation for the three most competitive L2R algorithms in the RankLib library and the set of features, i.e., baseline POP, baseline SAME, and using the SLP-feature. The columns depict the difference between recommending a property or RDF types for resources at subject or object positions of a triple. The overall recommendation quality of a L2R algorithm with a specific set of features is illustrated in the two most right columns

		sts		ps		ots		overall	
Model	Featues	MAP	MRR@5	MAP	MRR@5	MAP	MRR@5	MAP	MRR@5
CoordinateAscent	POP	.22 (.18)	.37 (.23)	.31 (.14)	.31 (.14)	.37 (.15)	.43 (.13)	.30 (.16)	.37 (.17)
	SAME	.26 (.16)	.33 (.24)	.29 (.13)	.29 (.13)	.39 (.13)	.47 (.12)	.31 (.14)	.36 (.16)
	SLP	.25 (.23)	.43 (.21)	.58 (.18)	.60 (.19)	.45 (.17)	.63 (.14)	.43 (.19)	.55 (.18)
LambdaMART	POP	.48 (.27)	.54 (.33)	.38 (.28)	.39 (.27)	.43 (.24)	.48 (.16)	.43 (.26)	.47 (.25)
	SAME	.48 (.26)	.57 (.29)	.40 (.26)	.40 (.26)	.41 (.23)	.51 (.19)	.43 (.25)	.49 (.25)
	SLP	.49 (.27)	.56 (.27)	.63 (.23)	.63 (.24)	.58 (.20)	.56 (.21)	.57 (.23)	.58 (.24)
RandomForests	POP	.44 (.29)	.55 (.31)	.35 (.28)	.36 (.28)	.43 (.25)	.49 (.26)	.41 (.27)	.47 (.28)
	SAME	.59 (.27)	.65 (.24)	.46 (.24)	.46 (.24)	.49 (.21)	.52 (.21)	.51 (.24)	.54 (.23)
	SLP	.65 (.26)	.70 (.24)	.63 (.25)	.63 (.24)	.64 (.17)	.68 (.15)	.64 (.23)	.67 (.21)

tices [18] recommend to reuse terms from popular vocabularies, therefore it is very likely that the ontology engineer initially has reused terms from popular vocabularies. This leads to a general coherence, which is trained by the Learning To Rank algorithm, that a vocabulary term from a popular vocabulary is likely to be a relevant recommendation candidate.

Recommendations based on reusing vocabulary terms from the same vocabulary (baseline SAME) have an MAP value of MAP $\approx .43$. A Friedman test ($\mathcal{X}^2 = 51,667, p < .001$) and the following Wilcoxon signed-rank test ($Z = -1.692, p = .011$) indicate that this difference in the recommendation quality is still significant compared to the baseline POP. However, it seems interesting that using the same-vocabulary-feature provides only an 8% gain in the absolute recommendation quality. Investigating the vocabulary terms used in the query-SLPs showed that many query-SLPs contain quite popular vocabulary terms, but they are rarely from the same vocabulary. In total, in 43% of the SLPs in the training and test set contained two or more terms from the same vocabulary. That means: the vocabulary terms that are extracted from an SLP before providing TermPicker with the resulting query-SLP are rarely from the same vocabulary as the remaining terms in the query-SLP. Thus, the L2R algorithms are less likely to regard this feature to provide more appropriate recommendations.

Using the SLP-feature increases the average MAP value up to MAP $\approx .70$. A Wilcoxon signed-rank test showed that using the SLP-feature and comparing its recommendation quality to the one of the baseline SAME, the p-value is $Z = -4.782, p < 0.001$. Due to the transitivity of this relation, the recommendation quality when using the SLP-feature is also significantly higher to the recommendation quality when using solely features to define popular vocabulary terms (baseline POP). Such a result depict to which extend the SLP-feature is relevant for providing valuable vocabulary term recommendations. Yet again, these results are based on using real-life data for calculating the query-SLPs for the evaluation. If the recommendation quality using the SLP-feature is that large, one can argue that the utilized real-life data was initially modeled by investigating which vocabulary terms other data providers have used to model their data. However, as establishing an ontological agreement in data representation is one central goal when reusing vocabularies [18], the results indicate that using recommendation based on the SLP-feature will eventually result in such a goal.

The evaluation based on the BTC 2014 data provides a more noticeable gain in the recommendation quality when using the SLP-feature than the evaluation based on the DyLDO data. In general, the key aspect of providing valuable recommendations lies in training the ranking model using representative data. In our

case, this includes the query-SLPs that are used to train the ranking model, but also the data that is used to calculate the five feature values for each recommendation candidate. Further investigations have shown, that the feature values calculated based on DyLDO data were less expressive compared to the feature values calculated based on the BTC 2014 dataset. In other words, the evaluation based on the BTC 2014 dataset provided an SLP-feature value of $f_5 > 0$ for 37% more relevant recommendation candidates than using the DyLDO data. The ranking model, which was learned based on the BTC 2014 data, therefore ranked recommendations with an SLP-feature value greater than zero rather to the top of the result list. This observation is validated by using the ranking models learned using the BTC 2014 data to rank the recommendation candidates for query-SLPs calculated from the DyLDO data. The resulting recommendation quality was $15\% - 20\%$ higher than using a ranking model learned based on the DyLDO data. The reason for such a difference in the recommendation quality is very likely the number of SLPs in the set $\mathbb{SLP}$, i.e., the SLPs that are calculated from existing datasets on the LOD cloud. Using the BTC 2014 dataset the number of such SLPs is twice as high compared to the number of such SLPs using the DyLDO data. As it is much more likely to calculate an SLP-feature value of $f_5 > 0$ with more SLPs contained in $\mathbb{SLP}$, it is quite reasonable that the evaluation based on the BTC 2014 data provides a higher recommendation quality.

6.1.3 RDF type recommendations vs. property recommendations (iii)

The differences between recommending RDF types a properties represent the different *modeling steps* in the engineering process of a schema [27]. It is accustomed to define a set of classes, which depict the entities that one wants to model, first, and then define relationships connecting these classes. Thus, it could be also accustomed that TermPicker recommends RDF types to describe the defined classes before recommending properties to interlink the RDF types. However, the differences in the recommendation quality between recommending RDF types for resources in subject or object position, or recommending properties seem to be marginal and cannot be considered significant according to the Friedman test, $\mathcal{X}^2 = 14,000, p = .449$ **n.s.**.

One aspect might be that the recommendation quality depends on how many vocabulary terms are already included in the query-SLP. In other words, a query-SLP containing three or more vocabulary terms could provide more concrete recommendations, than a query-SLP containing solely one term. For example, one would assume, that on one hand the query SLP slp_{q_1} with

$$slp_{q_1} = (\{\text{foaf:Person}\}, \varnothing, \{\text{foaf:Image}\})$$

produces more specific recommendation, due to the restriction of already reusing foaf:Person and foaf:Image. On the other hand a query-SLP slp_{q_2}, such as

$$slp_{q_2} = (\{\text{foaf:Person}\}, \varnothing, \varnothing)$$

should produce a larger amount of recommendations, as the query is not as restricted as the query slp_{q_1}. The chances of ranking a relevant vocabulary term to the top of the result should thus be higher for a query-SLP such as slp_{q_1}, i.e., query-SLPs that contain more vocabulary terms, as there is not as much noise in the recommendations. However, the differences between the query-SLPs with varying amount of contained vocabulary terms did not prove to be significant, $\mathcal{X}^2 = 15,800, p = .327$ **n.s.**. Therefore, one can conclude that TermPicker provides appropriate vocabulary term recommendations regardless if one is searching for RDF types describing resources in subject or object position of a triple, or for properties connecting two sets of RDF types. If another dataset on the LOD cloud uses a vocabulary term in conjunction with the terms included in the query-SLP, it has a large chance to be ranked at the top of the recommendation list.

6.2 Discussion of the Proposed Approach and the Evaluation

Learning to Rank tries to establish a correlation between the feature values of a recommendation candidate and its relevance [23]. Using the SLP-feature provides valuable results in most cases, but in the end the ranked results lists depend on the ranking model. Whether or not the SLP-feature is useful thus depends on the utilized training data, as demonstrated by the differences of using the BTC 2014 and the DyLDO data. For DyLDO, it does not work as well and leads to a decrease of the influence of the SLP-feature. This is because it does not contain a large variety of vocabulary terms and thereby decreases the chance of finding a term that has been used by other datasets on the LOD cloud in a similar way. The same applies for the same-vocabulary-feature. Generally, the proposed recommendation approach is reproducible with each Linked Data collection, e.g., with the BTC 2012 or the *Timbl* dataset which seed list contains URIs from Tim Berners-Lee's FOAF profile, but the bigger the data, the better the training data and the resulting ranking model. The best option would be to use the data from all datasets on the LOD cloud. However, computing SLPs from such a massive data collection is very time consuming and was not feasible for the provided evaluation.

The problem of finding an appropriate vocabulary term is a typical information retrieval problem that can be addressed via a machine learning approach. Thus, we validated the usefulness of the SLP-feature by using Learning To Rank, as it provides a methodology to induce a ranking model, that can be applied in general situations to retrieve appropriate vocabulary terms for reuse. Other approaches such as the Data Mining approach *Association Rules* conquer this problem by recommending terms based on the simple statement: "Datasets on the LOD cloud, who have used these vocabulary terms, have also used the following:...". This way, a vocabulary term that is not used in a similar manner will not be recommended. However, it also increases the chances that the result lists return empty. Therefore, it is rather a question whether the user also wants to get recommendation that make him/her "think outside the box", or whether he/she likes to stay as conform as possible to what others have used.

A potential threat to the validity of our experiments is the utilized evaluation design. It considers solely the recommendation candidates as relevant that have been extracted from a query-SLP before providing this query-SLP as input for TermPicker (cf. Section 4.1). This leads to two major vulnerabilities considering the validity of the evaluation. For once, many recommendation candidates are identified as irrelevant, although they are appropriate considering the rdfs:domain and rdfs:range, the owl:equivalentClass, or other information. For example, for the query-SLP slp_q with

$$\begin{aligned} slp_j &= (\{\text{swrc:Publication}\}, \{\text{swrc:author}\}, \{\text{foaf:Person}\}) \\ slp_q &= slp_j \ominus_{ps} \text{swrc:author} \\ &= (\{\text{swrc:Publication}\}, \varnothing, \{\text{foaf:Person}\}) \end{aligned}$$

the only relevant recommendation candidate for properties is swrc:author, as it was originally used. Properties, such as dc:creator or foaf:maker are considered as irrelevant in our evaluation, although it would make sense to reuse these properties to interlink resources of type swrc:Publication with resources of type foaf:Person. Thus, an L2R algorithm may identify many *appropriate* vocabulary terms (with an SLP-feature greater than zero) as irrelevant, which then can result in a ill-trained ranking model. Second, using many SLPs such as slp_j in the previous example, will favor point-wise L2R algorithms, as they tend to perform better, if there is only one or a few relevant items [23]. The previous example also shows, that there might be more than only a few relevant vocabulary terms. Utilizing a bigger set of relevant recommendation candidates might change the quality of point-wise, pair-wise, and list-wise L2R algorithms, such that list-wise and pair-wise algorithms might perform better than the point-wise approach. However, addressing this limitation requires human judgment whether a recommendation is relevant or not. Thus, conducting an experiment with human users is part of our future work.

7. RELATED WORK

The related work focuses on the schema-level patterns as well as on services that support an engineer in reusing vocabularies. The notion of schema-level pattern can be compared to the notion of so-called *triple patterns* [33], which essentially describe which property is in between a certain subject and a certain object. They can also be used to identify the RDF types of the subject and object, leading to the possibility of constructing a tuple that specifies which RDF type is connected to another type via a specific property. The tool for inspecting and exploring datasets Loupe,[18] makes use of these triple patterns to explore the triples in a dataset. Such result can also be achieved using a SPARQL query that retrieves the RDF types of a subject and an object as well as the connecting property between the subject and the object. However, both of these approaches contain solely one RDF type for the subject resource, one RDF type for the object resource, and one property connecting the resources. SLPs on the contrary may include more vocabulary terms to specify an RDF type of a resource or a property specifying a connection. It is a more condensed form of representation of the triple patterns and makes it easier to understand the data and faster to compute vocabulary terms recommendation. For example, the single SLP

$$(\{\text{foaf:Person}, \text{dbo:SoccerPlayer}\}, \{\text{foaf:knows}, \text{schema:colleague}\}, \{\text{schema:Person}, \text{dbo:Coach}\})$$

is enough to specify that resources of RDF types foaf:Person and dbo:SoccerPlayer are connected to resources of types schema:Person and dbo:Coach via the properties foaf:knows and schema:colleague. One would need eight triple patterns, i.e., every combination between the RDF types and the two properties, in order to specify the relationship. With each additional vocabulary term, the number of triples patters needed to represent the relationship rises drastically, such that it makes it harder to understand the data as well as more complicated to calculate recommendations from it.

7.1 Vocabulary Search Engines

Services providing a search for specific vocabulary terms generally utilize a keyword-based approach. Their input is a string describing the desired vocabulary term, e.g., a search-string "Person" to find vocabulary terms describing a person. The output is a set of RDF types and/or properties that are similar to the search-string based on some string similarity measure. Prominent and inspiring examples of such search engines are Swoogle [13], vocab.cc [31], Watson [12], Falcon's concept and object Search [8, 10], and LOV [34]. Falcons contains RDF types and properties from over $4,000$ ontologies, Swoogle even from over $10,000$ ontologies, whereas LOV comprises over 500 established manually curated vocabularies, and vocab.cc provides lists of the top 100 RDF types and properties in the Billion Triple Challenge 2011 data set. Each service provides also various meta-information on the vocabulary terms and their vocabularies, such as the term's number of usages on the LOD cloud. LOV also offers an API[19] that enables retrieving vocabulary terms by providing a query (e.g. "Person") and various other parameters, such as a *type* (e.g. "class") or even a *tag* specifying a category for a term (e.g."people"). The results are ordered based on a sophisticated ranking method adapting term frequency inverse document frequency (tf-idf) to the graph-structure of vocabularies [35]. Falcons concept search recommends further ontologies once the user selects an RDF type or a property from the result list, which can be investigated for further vocabulary terms. Falcons object search as well as Watson let the user search for specific entities, such as "Barack Obama", and retrieve resources from datasets on the LOD cloud that have properties containing the search string. This way, they are able to suggest RDF types and properties for the retrieved resources.

Alani et al. [2] propose another approach for searching ontologies from different domains. When searching for ontologies of a particular domain, a collection of terms that represent the given domain is retrieved and used to expand the user query. This is especially helpful when starting to choose vocabulary terms for reuse from scratch.

As additional information on the vocabulary terms, most services exploit schema-information encoded in the vocabularies, such as sub-class, sub-property or other relations between vocabulary terms. For example, LOV offers a SPARQL[20] endpoint for Linked Data practitioners and applications. Using the endpoint, one can search for RDF types or properties that are equivalent or a sub-class-of another vocabulary term, or one can search for properties that have a specific RDF type as rdfs:domain and/or rdfs:range. Listing 2 illustrates an example query to retrieve RDF types from all vocabularies stored in LOV that are equivalent to foaf:Person. SPARQL queries for selecting RDF types that are a rdfs:subClassOf another RDF type, or properties having a specific domain and range can be designed analogously. The results of such queries however de-

```
1 PREFIX owl: <http://www.w3.org/2002/07/owl#>
2 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
3
4 SELECT DISTINCT ?t {
5   GRAPH ?src{
6     ?t owl:equivalentClass foaf:Person.
7 }} ORDER BY ?t
```

Listing 2: SPARQL query in LOV. Querying for RDF types (`?t`) from all vocabularies/graphs in LOV (`?src`) that are equivalent to the RDF type foaf:Person. This enables to exploit structural information encoded in the RDF vocabularies

[18] `http://loupe.linkeddata.es/loupe/`, last access 12/12/15

[19] `http://lov.okfn.org/dataset/lov/api`, last access 09/03/2015

[20] `http://lov.okfn.org/dataset/lov/sparql`, last access 09/03/2015

pend on whether the connections are defined in the vocabularies stored, i.e., T-Box specification of vocabularies. Equivalent vocabulary terms cannot be retrieved, if vocabulary terms are not connected via links such as owl:equivalentClass, rdfs:subClassOf, owl:equivalentProperty, or others.

However, such additional information on vocabulary terms do not depict how other datasets on the LOD cloud describe their data. TermPicker does not rely on T-Box information and the modeled connections between vocabularies, nor does it suggest vocabulary terms for *specific* resources. It rather uses the datasets on the LOD cloud to calculate schema-level patterns representing which vocabulary terms are used to describe all resources and their connections. This way, TermPicker's recommendations are based on the encoded A-Box specification of datasets published on the LOD cloud. On contrary to Falcons' object search and Watson, TermPicker's input is a set of vocabulary terms, i.e., a query-SLP. Using this input, it generates a list of other vocabulary terms that are used in conjunction with the sets contained in the query-SLP, and not a list of vocabulary terms used to describe one specific resource. Based on this, TermPicker is able to retrieve further RDF types to describe resources of a given RDF type. For example, for resources of type foaf:Person it can suggest the RDF types schema:Person, swrc:Person, dbp:FootballPlayer, if other LOD data providers have used these types in combination with foaf:Person to describe their data. Summarizing, with TermPicker, connections between vocabulary terms do not have to be made explicit, but can be induced directly from datasets on the LOD cloud.

7.2 Vocabulary Recommender Systems

Existing services that recommend RDF types and properties are generally based on syntactic and semantic similarity measures as well as on algorithms that provide a statement on the popularity of a recommended term. One prominent example is the collaborative system CORE for ontology engineering (short for: Collaborative Ontology Reuse and Evaluation) [15]. A set of initial keywords defines CORE's input. Starting from this, CORE determines a ranked list of domain-specific ontologies considerable for reuse. The approach uses WordNet[21] to expand the initial set of terms, and performs a search for each of the defined keywords on an index of ontologies. Besides syntactic and semantic similarity measures, CORE uses manual user evaluations of suggested ontologies to raise the recommendation quality. A similar system was developed by Romero et al. [28]. However, it measures the popularity of an ontology by the number of appearances in Wikipedia or bookmarks on Del.icio.us[22]. The previously mentioned service Watson [12] also provides a plug-in for the NeOn ontology engineering toolkit[23] that supports the engineer in reusing vocabularies. It uses semantic information from a number of ontologies and other semantic documents published on the Web to recommend appropriate vocabulary terms.

Again, the input for these recommendation services is a single string or a set of strings specifying a vocabulary term or a domain of interest. Whereas these services provide recommendations based on string analyzes, they do not exploit any structural information on how vocabulary terms are connected to each other. In contrast, Falcons' Ontology Search [9] provides the engineer with such information. Compared to traditional ontology matching approach, which align ontologies based on *similarity*, the authors of Falcons' Ontology Search use different kinds of relatedness, in order to identify which vocabulary terms might express similar semantics. However, it is mainly designed to establish a general relatedness between vocabularies specifying that different vocabularies contain terms that describe similar data. Thus, it does not investigate how data providers on the LOD cloud use vocabulary terms to describe their data and individual relations as it is done by TermPicker.

There are various tools and services that transform data from various formats, such as CSV data, into RDF, but only a few ones provide support for reusing vocabularies by integrating a vocabulary recommendation service. The "data2Ontology" module of the Datalift platform [30] provides suggestions to match data entities to a vocabulary term based linguistic proximity between the data entity and the vocabulary term and the quality of the vocabulary using criteria from LOV. The data integration tool Karma [21] contains two different types of recommending vocabulary terms. One approach suggests so-called *semantic types* for a column containing data, such as the first name of a person [22]. The approach analyzes the content of the column using NLP techniques and recommends an RDF type in conjunction with a datatype property containing the literal value of a column's cell. The other recommendation approach is based on what the user has previously modeled [32]. For example, if she has already modeled data entities and relationship about museum items, and the next data collection contains data on other museum items, the system is likely to recognize this and recommends the vocabulary terms that were used to model the previous data collection. However, these tools either use string similarity, analyze the modeled data entities themselves, or rely on previously modeled data by the user, and do not consider what other data providers on the LOD cloud have used to model their data. This way, their input for providing recommendations is very different and cannot be directly compared to TermPicker and its approach.

7.3 Ontology Matching and Alignment

As schema-level patterns can be used to describe Linked Open Data on schema-level, and given that TermPicker compares the query-SLP with other SLPs calculated from existing datasets on the LOD cloud, one might consider TermPicker's approach being related to ontology matching [14]. However, typical ontology matching techniques try to find correspondences between semantically related vocabulary terms of two or more different ontologies by applying (semi-)automatic alignment algorithms. In contrast, SLPs solely represent the connection between resources of specific RDF types via a set of properties. The comparison of two SLPs is done solely syntactically, i.e., if the two sets of RDF types and the set of properties of two SLPs contain the same vocabulary terms, these two SLPs are considered the same. Thus, SLPs do not find any correspondences between semantically related vocabulary terms and is therefore not some type of ontology matching technique, nor can it be directly compared to such.

8. CONCLUSION

This paper presented TermPicker: a novel approach for recommending vocabulary terms for reuse. The notion of schema-level patterns (SLPs), which are a major part of TermPicker, was introduced including the description how they are calculated from datasets on the LOD cloud. It has been demonstrated how SLPs are used to define the SLP-feature and how Learning To Rank algorithms use the features to train a ranking model. Two 10-fold leave-one-out evaluations were performed on the BTC 2014 and

[21] http://wordnet.princeton.edu/, last access 09/05/15
[22] http://delicious.com/, access 7/15/2014
[23] http://www.neon-project.org/, last access 09/05/15

the DyLDO dataset, respectively, and the results illustrate that using the SLP-feature provide vocabulary term recommendation with a Mean Average Precision of about MAP ≈ 0.70. This improves the recommendation quality by about 35% compared to the baselines of recommending vocabulary terms from popular vocabularies and recommending terms from the same vocabulary. Furthermore, with a Mean Reciprocal Rank at the first five positions of $MRR@5 \approx 0.74$, the results indicate that the first relevant vocabulary term recommendation is within the first five results in 74% of all queries. Finally, based on the evaluation design that assesses the relevance of a recommendation candidate automatically (by extracting some terms from a query-SLP before using it as input for TermPicker), it seems that point-wise Learning To Rank (L2R) algorithms provide better results than pair-wise or list-wise L2R algorithms.

As future work, we intend to compare the data mining approach *Association Rule Mining* to the utilized L2R algorithms. To do so, we compose a user-study, in which the user gets recommendations based on Learning To Rank or on Association Rules for a specific assignment. The user subsequently rates the perceived recommendation quality of both approaches, such that we can compare which approach provides the overall better recommendations.

To increase the recommendation quality based on Learning To Rank, it seems useful to consider the domain in which a vocabulary term is used most often. As another feature, one could use the PageRank information of a given pay-level domain that uses a recommended vocabulary term. This way, recommendations can be more domain specific. Furthermore, each recommendation candidate can be enriched with meta-information, such as the appropriate rdfs:domain and rdfs:range information for properties, or owl:equivalentClass or other information for RDF types. To this end, LOV provides an API or a SPARQL endpoint that can be used.

Acknowledgment. We thank the staff of ZWB Kiel for providing feedback for this research. Specifically, Norbert Luttenberger, who, inspired by the tool *ColumnPicker*, proposed the name *TermPicker*, as it resembles a similar useability for picking vocabulary terms instead of columns. Furthermore, Jesper Zedlitz provided valuable insights on distinguishing the proposed approach to existing tools and services, using the the specifications of the T-Box and the A-Box.

9. REFERENCES

[1] A. Agarwal, H. Raghavan, K. Subbian, P. Melville, R. D. Lawrence, D. C. Gondek, and J. Fan. Learning to rank for robust question answering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 833–842. ACM, 2012.

[2] H. Alani, N. F. Noy, N. Shah, N. Shadbolt, and M. A. Musen. Searching ontologies based on content: experiments in the biomedical domain. In *K-CAP*, pages 55–62. ACM, 2007.

[3] S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats–an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management*, pages 353–362. Springer, 2012.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96. ACM, 2005.

[6] R. Busa-Fekete, G. Szarvas, T. Elteto, B. Kégl, et al. An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In *ECAI 2012-20th European Conference on Artificial Intelligence*, volume 242, 2012.

[7] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

[8] G. Cheng, W. Ge, and Y. Qu. Falcons: searching and browsing entities on the semantic web. In *Proceedings of the 17th international conference on World Wide Web*, pages 1101–1102. ACM, 2008.

[9] G. Cheng, S. Gong, and Y. Qu. An empirical study of vocabulary relatedness and its application to recommender systems. In *The Semantic Web–ISWC 2011*, pages 98–113. Springer, 2011.

[10] G. Cheng and Y. Qu. Searching linked objects with falcons: Approach, implementation and evaluation. pages 49–70, 2009.

[11] N. Craswell. Mean reciprocal rank. In *Encyc. of Database Systems*. Springer, 2009.

[12] M. d'Aquin, C. Baldassarre, L. Gridinoc, M. Sabou, S. Angeletou, and E. Motta. Watson: Supporting next generation semantic web applications. 2007.

[13] L. Ding, T. W. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM*, pages 652–659, 2004.

[14] J. Euzenat, P. Shvaiko, et al. *Ontology matching*, volume 18. Springer, 2007.

[15] M. Fernandez, I. Cantador, and P. Castells. Core: a tool for collaborative ontology reuse and evaluation. In *4th International Workshop on Evaluation of Ontologies for the Web*, 2006.

[16] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[17] L. Hang. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

[18] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.

[19] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data*, pages 213–227. Springer, 2013.

[20] T. Käfer and A. Harth. Billion Triples Challenge data set. Downloaded from http://km.aifb.kit.edu/projects/btc-2014/, 2014.

[21] C. Knoblock, P. Szekely, J. Ambite, A. Goel, S. Gupta, and et al. Semi-automatically mapping structured sources into the semantic web. In *The Semantic Web: Research and Applications*, volume 7295, pages 375–390. Springer Berlin/Heidelberg, 2012.

[22] R. Krishnamurthy, A. Mittal, C. A. Knoblock, and P. Szekely. Assigning Semantic Labels to Data Sources. In *Proceedings of the 12th Extended Semantic Web Conference (ESWC)*, May 2015.

[23] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[24] G. Lodi, A. Maccioni, M. Scannapieco, M. Scanu, and L. Tosco. Publishing official classifications in linked open data. 2014.

[25] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[26] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[27] N. F. Noy and D. L. McGuinness. *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, 2001.

[28] M. M. Romero, J. M. Vázquez-Naya, C. R. Munteanu, J. Pereira, and A. Pazos. An approach for the automatic recommendation of ontologies using collaborative knowledge. In *KES (2)*, volume 6277 of *LNCS*, pages 74–81. Springer, 2010.

[29] J. Schaible, T. Gottron, and A. Scherp. Survey on common strategies of vocabulary reuse in linked open data modeling. In *ESWC*. Springer, 2014.

[30] F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, and et al. Enabling linked-data publication with the datalift platform. In *AAAI 2012, 26th Conf. on Artificial Intelligence - Semantic Cities*, 2012.

[31] S. Stadtmüller, A. Harth, and M. Grobelnik. Accessing information about linked data vocabularies with vocab.cc. In *Semantic Web and Web Science*, pages 391–396. Springer, 2013.

[32] M. Taheriyan, C. Knoblock, P. Szekely, and J. L. Ambite. Learning the Semantics of Structured Data Sources. *Journal of Web Semantics Special Issue on Knowledge Graphs*, 2016.

[33] T. Tran and G. Ladwig. Structure index for rdf data. In *Workshop on Semantic Data Management (SemData@ VLDB)*, volume 2, 2010.

[34] P.-Y. Vandenbussche, B. Vatant, and L. Rozat. Linked open vocabularies: an initiative for the web of data. In *QetR Workshop, Chambery, France, 2011*, 2011.

[35] P.-Y. Vandenbusschea, G. A. Atemezingb, M. Poveda-Villalónc, and B. Vatantd. Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. *Semantic Web Journal*, 2015.

[36] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[37] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *SIGIR*, pages 391–398. ACM, 2007.