VocRec: An Automated Vocabulary Recommender Tool

Wagner G. do Amaral[1], Bernardo Pereira Nunes[1,2], Sean W. M. Siqueira[1], Luiz André P. Paes Leme[3]

[1] Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro/RJ, Brazil
{wagner.amaral,bernardo.nunes, sean}@uniriotec.br
[2] Department of Informatics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro/RJ, Brazil
bnunes@inf.puc-rio.br
3 Institute of Computing, Fluminense Federal University (UFF), Niterói/RJ, Brazil
lapaesleme@ic.uff.br

Abstract. A common problem faced by data publishers is how to semantically represent data and how to find vocabularies that best represent their data to publish in the Web. To address these problems, this paper introduces a tool, called VocRec, which can be used for recommending vocabularies for relational data that will be published as Linked Data on the Web. Preliminary evaluation using educational databases shows promising results in terms of precision and its ability on assisting data publishers in the task of dataset publication.

Keywords: Vocabulary Recommendation, Linked Data.

1 Introduction

With the increasing adoption of Linked Data (LD) standards for publishing and connecting structured data on the Web, a global data space has been created covering a large variety of domains (e.g. Government, Life Sciences, Linguistics, etc.) as shown in the LOD cloud [1]. Additionally, LD has also led to the creation of a number of cross-domain and domain-specific applications [2,3], allowing the interlinking of heterogeneous applications at the data level.

Despite the many benefits of using LD, a number of challenges arises when one wants to publish data following the LD standards. A common problem faced by data publishers is how to semantically represent data – one of the first steps when publishing linked data [4]. For this task, a data publisher needs to either create his own vocabulary or to reuse one or more of the existing vocabularies published on the Web. The latter takes data publishers to a prior problem, that is, how to find vocabularies that best represent their relational data? This is the problem addressed in this paper.

Vocabularies are responsible for adding semantics to data, defining and characterizing concepts, relationships and constraints [5]. The use of well-known and largely adopted vocabularies is key to enable data integration and interoperability [6]. Following the recommendation provided by [7], a new vocabulary should only be defined if there is no other mix of existing vocabularies that can represent your data. However, there exists a thousand of vocabularies [8] representing multiple areas of concern and very often a single vocabulary is not sufficient to entirely represent a dataset. For instance, suppose that a data publisher wants to publish a relational database with hundreds of entities following the LD standards. The mapping between the entities and the most adequate vocabularies would be time-consuming and very hard even for specialists.

This paper introduces a tool named VocRec, an automated vocabulary recommender tool, to assist data publishers in selecting vocabularies to represent their relational data. The tool follows a 5-step process chain starting from the extraction of syntactical information from relational databases until the actually vocabulary recommendation after passing through clustering and a semantic processing. A preliminary evaluation shows promising results on the recommendation of vocabularies.

2 VocRec – Automated Vocabulary Recommender Tool

This section introduces the VocRec tool through a running example. We use a database schema from a well-known learning management system called Moodle [9] to exemplify the whole recommendation process. Its database schema contains a total of 2,840 attributes scattered into 314 tables. The choice of this schema is motivated by its complexity and the diverse knowledge it represents.

Figure 1 illustrates the tool and overviews the process chain responsible for recommending vocabularies to data publishers. The process is split into five main steps, namely: (i) Data Extraction; (ii) Data Preprocessing; (iii) Contextual Clustering; (iv) Semantic Expansion; and (v) Vocabulary Recommendation. In what follows we describe the vocabulary recommendation process chain and instantiate each step based in Moodle's relational database used as input to the process:

(i) Data Extraction. This step is responsible for the extraction of schema information [10] from relational database (e.g. tables, attributes and relationships).

Given the Moodle schema, this step outputs the names of tables (e.g., *mdl_user, mdl_course_categories)* and attributes (e.g., *fullname* and *course_description).* After extracting schema information, a preprocessing step is required to clean up messy data.

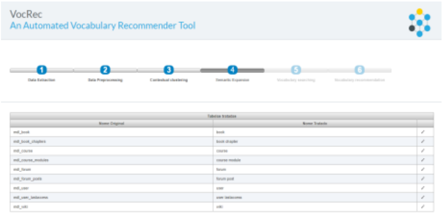
(ii) Data Preprocessing. This step is responsible for cleaning and preparing the data for the next steps. For this, Apache OpenNLP toolkit[1] is used to detect the language, tokenize, and stemming.

Continuing with the running example, after data preprocessing, the entities *mdl_user*, *mdl_course_categories* are represented by [*user*] and [*course, category*]. Note that "*mdl_*" is a prefix and is removed during the preprocessing, and *categories* is now in its singular form after stemming.

(iii) Contextual Clustering. This step is responsible for grouping related entities for further recommendation of vocabularies. There are at least two possible ways to

[1] https://opennlp.apache.org/

group related entities: (a) using the existing relationships (e.g. *foreign keys)* between entities in a database; and/or (b) using (semantic) similarity between terms. VocRec implements the latter approach. For this, we use semantic similarity and relatedness measures from the lexical database WordNet through the API WS4J[2], and based on the similarity of the terms the clusters are formed using the K-Means algorithm.



Tabelas tratadas		
Nome Original	Nome Tratado	
mdl_book	book	✎
mdl_book_chapters	book chapter	✎
mdl_course	course	✎
mdl_course_modules	course module	✎
mdl_forum	forum	✎
mdl_forum_posts	forum post	✎
mdl_user	user	✎
mdl_user_lastaccess	user lastaccess	✎
mdl_wiki	wiki	✎

Fig. 1. The automated processing chain of VocRec to recommending vocabularies.

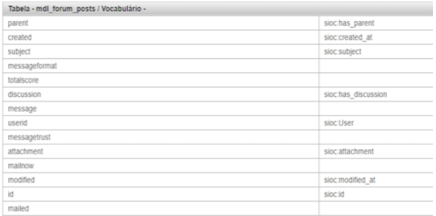


Tabela - mdl_forum_posts / Vocabulário -	
parent	sioc:has_parent
created	sioc:created_at
subject	sioc:subject
messageformat	
totalscore	
discussion	sioc:has_discussion
message	
userid	sioc:User
messagetrust	
attachment	sioc:attachment
mailnow	
modified	sioc:modified_at
id	sioc:id
mailed	

Fig. 2. Vocabulary recommendation for the table *mdl_forum_posts*. The recommended vocabulary is SIOC (Semantically-Interlinked Online Communities).

For instance, assume that the previous steps extracted information about the entities: *mdl_user*, *mdl_course_modules* and *mdl_course_category*. So, after preprocessing, the following vectors are generated: [*user*], [*course, module*] and [*course, category*]. Clearly, two clusters should be generated based on the semantic similarity between the vectors: Cluster 1: {[*user*]}, and Cluster 2: {[*course, module*], [*course, category*]}.

(iv) Semantic Expansion. This step is responsible for automatically extracting semantic relationships from the clusters generated in the previous step. Here, we use the MIT Java Wordnet Interface to synsets and sense index from Wordnet based on the terms representing the clusters. A cluster is represented by the most occurring terms, so, for instance, the term representing Cluster 2 is *course*. Based on the synsets and the sense of the term *course*, we expand the cluster representative terms by using their synonyms and hypernyms. For example, the term *user* found in Cluster 1 has as its hypernym the term “Person”, which can be used to find related vocabularies in the next and final step. So, the output of this step is a set of related terms that represent a group of entities extracted from a relational database schema.

(v) Vocabulary Recommendation. This step is responsible for recommending vocabularies. Based on the set of terms generated in the previous step, the terms are used to query a vocabulary repository. Two queries are issued, one containing the hypernym term as a class, and another with all expanded terms. The vocabulary repository used is LOV. Three different strategies are available to the vocabulary recommendations: (i) Popular Vocabularies; (ii) Minimize the number of vocabularies; and (iii) Maximize the number of vocabularies. The first strategy takes into account the usage of the vocabularies in order to recommend the most popular ones (thus, the most probable to be used) whereas the second and the third strategies use the mini-

[2] https://github.com/Sciss/ws4j

mum/maximum number of vocabularies to represent the data. Figure 2 shows the recommended vocabulary for the entity *mdl_forum_posts* using the *popular* strategy.

3 Preliminary Evaluation and Results

The preliminary evaluation of the quality of the results obtained were conducted based on precision and used databases from the Education field. Table 1 summarizes the results. Although, on average, the best strategy is based in the popular vocabularies, the other two strategies may show to data publishers other vocabularies that may also represent the data.

Table 1. Table results the vocabulary recommendation to two groups of databases.

DB schema	**Popular Vocabularies**			**Minimize vocabularies**			**Maximize vocabularies**		
	Moodle	Sakai	Atutor	Moodle	Sakai	Atutor	Moodle	Sakai	Atutor
Precision	0.58	0.70	0.72	0.34	0.46	0.66	0.27	0.40	0.78

4 Conclusion and Future Works

This paper introduced VocRec a tool used for recommending vocabularies for relational data that will be published as LD in the Web. Preliminary evaluation shows promising results and its ability on assisting data publishers in the task of dataset publication. The recommendation of vocabularies for Atutor reached 78% of precision using the maximize strategy whereas the precision for Moodle and Saki reached 58% and 70%, respectively, using the popular strategy. As future works, we intend to test VocRec in other scenarios other than Educational, perform in-depth evaluation and include an additional step to perform property alignment. We also intend to evaluate the usability of the tool. VocRec is publicly available at: https://tinyurl.com/y8m5pvtv

References

1. Cyganiak, R., Jentzsch, A.: The Linking Open Data cloud diagram, http://lod-cloud.net/, last accessed 2017/07/15.
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. 1, 1–136 (2011).
3. Mouromtsev, D., DAquin, M.: Open data for education linked, shared, and reusable data for teaching and learning. Springer, Switzerland (2016).
4. Best Practices for Publishing Linked Data, https://www.w3.org/TR/ld-bp/, last accessed 2017/07/15.
5. W3C, https://www.w3.org/standards/semanticweb/ontology, last accessed 2017/07/15.
6. Data on the Web Best Practices, https://www.w3.org/TR/dwbp/, last accessed 2017/07/15.

7. Bizer, C., Cyganiak, R., Heath, T.: How to Publish Linked Data on the Web, http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/#whichvocabs, last accessed 2017/07/15.
8. Vandenbussche, P.-Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. Semantic Web. 8, 437–452 (2017).
9. Moodle - Modular object-oriented dynamic learning environment, https://moodle.org/, last accessed 2017/07/15.
10. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal. 10, 334–350 (2001).