Information Content based Ranking Metric for Linked Open Vocabularies

Ghislain Auguste Atemezing
EURECOM
Campus SophiaTech, France
atemezin@eurecom.fr

Raphaël Troncy
EURECOM
Campus SophiaTech, France
raphael.troncy@eurecom.fr

ABSTRACT

It is widely accepted that by controlling metadata, it is easier to publish high quality data on the web. Metadata, in the context of Linked Data, refers to vocabularies and ontologies used for describing data. With more and more data published on the web, the need for reusing controlled taxonomies and vocabularies is becoming more and more a necessity. Catalogues of vocabularies are generally a starting point to search for vocabularies based on search terms. Some recent studies recommend that it is better to reuse terms from "popular" vocabularies [3]. However, there is not yet an agreement on what makes a popular vocabulary since it depends on diverse criteria such as the number of properties, the number of datasets using part or the whole vocabulary, etc. In this paper, we propose a method for ranking vocabularies based on an information content metric which combines three features: (i) the datasets using the vocabulary, (ii) the outlinks from the vocabulary and (iii) the inlinks to the vocabulary. We applied this method to 366 vocabularies described in the LOV catalogue. The results are then compared with other catalogues which provide alternative rankings.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.5 [**Online Information Services**]: Data sharing—*Web-based services*

General Terms

Ranking, Linked Data, Vocabularies

Keywords

Information Content, Linked Open Vocabularies, reusing vocabularies, ranking metric.

1. INTRODUCTION

The linked data principles have gained significant momentum over the last few years as a best practice for sharing and publishing structured data on the Semantic Web [1]. Before being published, data is modeled and ontologies or vocabularies are one of the key elements of a dataset. Vocabularies are the artefact that bring semantics to raw data. One of the major barriers to the deployment of linked data is the difficulty for data publishers to determine which vocabularies should be used since developing new vocabularies has a cost. Catalogues of ontologies are therefore a useful resource for searching terms (classes and properties) defined in those vocabularies. The Linked Open Vocabulary (LOV) initiative [7] is playing a significant role in providing such services to users who can search within curated vocabularies, fostering ontologies reuse. LOV focuses only on vocabularies submitted by users, which are then reviewed and validated by curators. In addition, LOV computes dependencies between vocabularies, keeps track of different versions of them in order to enable their temporal evolution.

To the best of our knowledge, recommending vocabularies to reuse are limited to "popular" or "well-known" ones. This paper proposes a metric combining different features such as how vocabularies are interlinked, or how they are used in real world datasets. This contribution originates also in the desire to bring the traditional concept of Information Content (IC) into the field of the semantic web applied to vocabularies. Many catalogs of ontologies already provide some ranking metrics based on some features. However, we are interested in applying the principles of IC on vocabularies to investigate if such techniques can give more insights in ontology ranking and ontology usage (e.g in visualization applications).

The paper is organized as follows: Section 2 defines the theory of Information Content, and the features used for applying Partition Information Content to vocabularies. We present our experiments on the Linked Open Vocabularies catalogue in the Section 3. We discuss how this ranking metric can be used for vocabulary design and maintenance in Section 4. We compare our results with other rankings for vocabularies in Section 5 before concluding and outlining future work (Section 6).

2. INFORMATION CONTENT METRICS

Based on probability theory, Information Content (IC) is computed as a measure of generated amount of surprise [5]. More common terms in a given corpus with higher chance of occurrence cause less surprise and accordingly carry less information, whereas infrequent ones are more informative.

SEM '14, September 04 - 05 2014, Leipzig, AA, Germany

We reuse the notion of informativeness as the value of information associated with a given entity, where Information Content has a negative relation with its probability. The concept of Information Content can be used to rank each entity, term, or alphabet in the corpus. We apply the Partitioned Information Content to measure the informativeness of Linked Open Vocabularies as a semantic network of resources connected together using different range of relations, as described in [4]. Partitioned Information Content (PIC) is derived from the IC value using some weights. We empirically set those weights according to three features:

- (i) datasets using the vocabulary ($weight = 2$);
- (ii) *outlinks* from a vocabulary, i.e. whether a vocabulary reused other vocabularies ($weight = 1$);
- (iii) *inlinks* to a vocabulary, i.e. whether other vocabularies are reusing this vocabulary ($weight = 3$).

2.1 Information Content in Linked Open Vocabularies

This experiment aims at bringing the concept of informativeness in the field of terms semantically related as it is the case within semantic web ontologies. The ranking obtained can give additional information based on the Information Content theory to help reusing terms and detecting the ones that are less popular. This can then be used by applications consuming datasets described with these vocabularies. The equation (1) gives the formula for computing the IC value of a term (class or property):

$$IC(t) = -\log_2(\frac{\varphi(t)}{N}), \tag{1}$$

where N is set to be the maximum value corresponding to the term occurrence in the LOV aggregator (as of June 2014, this value is 3958, and it corresponds to the popularity of the `skos:prefLabel` property); and $\varphi(t)$ is the occurrence of the term (but not its popularity).

For computing $\varphi(t)$, we use two types of SPARQL queries depending on whether the term is a class (Listing 1) or a property (Listing 2 considers `owl:ObjectProperty`, `owl:DatatypeProperty` and `rdfs:Property`). Note that we do not yet take into account the `owl:equivalentClass` and `owl:equivalentProperty` axioms that may appear in some vocabularies. We leave this as a future work.

Listing 1: SPARQL query for computing the occurrence of a class

```
SELECT (count(?uri1) as ?occ)
WHERE {
   ?uri1  ?p %%classURI . }
```

2.2 Ranking Vocabularies using Information Content

For computing the PIC value, we use the following formula:

$$PIC(f) = w_f \times \sum_{i=1}^{n} IC(t_i), \tag{2}$$

Listing 2: SPARQL query for computing the occurrence of a property

```
SELECT (count(?uri1) as ?occ)
WHERE {
   ?uri1 +objectURI+ ?uri2.
   FILTER  (?uri1 != ?uri2) }
```

where w_f is the weight related to vocabulary f.

We consider very important that a vocabulary is being reused by other vocabularies and implemented within real world datasets. For example, the `foaf` ontology is weighted 6 because it reuses vocabularies (1), it has been used in some datasets (2) and it is being reused by other vocabularies (3). The `dul`[1] vocabulary is weighted 3 because it doesn't reuse any vocabulary but it is instead used by several other vocabularies.

3. EXPERIMENTS ON VOCABULARIES

We use the LOV catalogue, and particularly the LOV aggregator[2] to look at the terms (classes and properties) to compute their Information Content (IC). LOV defines the *LOV Distribution* as the number of vocabularies in LOV that refer to a particular element and the *LOV popularity* as the number of other vocabulary elements that refers to a particular one. Based on the concept of Partitioned Information Content, we implement our ranking measure according to the algorithm 1. We take the subset of classes and/or properties with LOV popularity and LOV distribution greater than one. The initial set of vocabularies in LOV is 366. After filtering the candidate terms, we came out with a set of 161 vocabularies (44% or 161 vocabularies) for computing their ranking.

The Table 1 gives the Top 15-ranking of the vocabularies according to the informativeness of the classes and properties used within the LOV ecosystem. As the function is proportional to the number of terms, we use a threshold of 22 terms in the vocabularies. For example, the PIC value of `dcterms` is higher than `foaf`'s because the former uses 53 terms (39 properties and 14 classes), while the latter only 35 terms (9 classes and 26 properties); although they both have the same weight value.

The Table 2 shows the Top 20 namespaces of vocabularies according to the informativeness of the classes and properties used within the LOV ecosystem, along with their Information Content Value.

4. APPLICATION OF INFORMATION CONTENT ON VOCABULARIES

We foresee various applications using the ranking method based on the Information Content metric while designing semantic web applications, vocabulary life-cycle management or novel recommendation services. We make the following

[1] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`

[2] `http://lov.okfn.org/endpoint/lov_aggregator`

Algorithm 1 Ranking vocabularies algorithm

```
Require: Dump of $lovaggregator$ file
Upload in a triple store for querying
Select subset of candidate vocabs $LOVaggregatorendpoint$
for $term \in lovaggregator$ do
    if ($LOVdistribution \geq 1$ )and ($LOVPopularity \geq 1$) then
        $candidateterms \leftarrow$ append $term$
    end if
end for
for each $term \in candidateterms$ do
    GROUP BY vocabulary namespace
    COMPUTE weight for each vocabulary
end for
INITIALIZE $PICvector$ AS a vector
for each $term \in candidateterms$ do
    while $term \in vocabularySpace$ do
        $ICterm \leftarrow$ function IC(term, vocabPrefix)
        $ICvocab \leftarrow \sum ICterm$
    end while
    $PICvocab \leftarrow weight(vocab) \times$ICvocab
    $PICvector \leftarrow$ append (PICvocab)
    ORDER PICvector
end for
return PICvector
```

recommendations when using the PIC ranking method on vocabularies:

- Vocabularies on the Top PIC-ranking can be used in visualization applications, i.e. to be displayed to the user as much as possible.
- Terms with lower IC can be used in facetted browsing, and they seem appropriate for generating `sameAs` links during the interconnection and enrichment process. They might also be used for promoting the reuse of terms in vocabularies in general.
- The PIC-ranking could help the ontology designers to monitor and to assess the usage of some terms and lead to update the ontology accordingly. For example, it can be useful in extending the use of the properties such as `vs:term_status` or `owl:deprecated`.
- Such a ranking can be used to rank organizations or publishers of vocabularies in a time period (e.g. annual) as a way to encourage good qualities vocabularies and/or datasets on the cloud.

The use of the information content on LOV vocabularies can be applied in the datasets interlinking task and visualization applications workflow. For interlinking datasets, this method can help detecting properties with a lower PIC which will be a candidate for the interlinking tool. The PIC score can further be used to track the vocabularies terms status (i.e. `vs:term_status`) or `owl:deprecated` properties by dataset maintainers. From the list of namespaces having deprecated terms (Table 3), we observe some correlations with the PIC rank for the vocabularies `dcat (8), vcard (36), gr (6), wl (2), pav (1)` and `bibo (1)`[3]. More precisely, the presence of `gr` and `bibo` provides evidence of such a correlation, while the presence of `dcat` and `card` can be explained by the fact that those two vocabularies are in a review process at W3C and subject to re-modeling respectively.

Rank	Prefix	PIC score
1	dcterms	1724.844
2	schema	1588.700
3	gr	1261.101
4	foaf	1033.197
5	bibo	876.205
6	time	816.2020
7	skos	805.287
8	dul	797.328
9	ptop	773.167
10	rdafrbr	640.834
11	vaem	630.621
12	ma-ont	508,694
13	prov	497.524
14	swrc	437.394
15	dce	428.618

Table 1: Top 15 vocabularies according to their PIC. All the prefixes used for the vocabularies are the ones used by LOV

Rank	vocab term	IC value
1	skos:example	7.7806
2	dce:contributor	4.674
3	skos:scopeNote	4.365
4	dcterms:source	4.299
5	mads:code	3.922
6	mads:authoritativeLabel	3.922
7	vs:userdocs	3.847
8	dce:title	3.79
9	skos:hasTopConcept	3.4547
10	dce:description	2.758
11	dcterms:issued	2.553
12	dce:creator	2.518
13	skos:inScheme	2.202
14	skos:notation	1.924
15	dcterms:description	1.646
16	coll:List	0.761
17	vs:term_status	0.735
18	skos:definition	0.43
19	skos:prefLabel	0.009
20	foaf:Person	0

Table 2: Ranking of Top 20 terms (classes and properties) according to their IC value

Table 3 gives an overview of some namespaces with their deprecated terms.

5. RELATED WORK AND DISCUSSION

In this section, we look at three other catalogues providing rankings for vocabularies: vocab.cc, LODStats and pre-

[3] As of June 2014, there are 60 terms deprecated in LOV with the query `http://bit.ly/1aqcDf3`

prefix	#DeprecatedTerms	dcterms:modified
vcard	36	2013-09-25
dcat	8	2013-09-20
gr	6	2011-10-01
wl	2	2013-05-30
pav	1	2013-08-30
bibo	1	2009-11-04

Table 3: Sample of vocabularies with terms deprecated in LOV

fix.cc. vocab.cc[4] does not provide a ranking for vocabularies but rather proposes a rank for classes and properties. The proposed ranking presented in Table 4 is taken from the ranking of classes assuming the namespace is used only once per class.

Rank	**LOV-PIC**	**prefix.cc**	**vocab.cc**	**lodstats**
1	dcterms	yago	intervals	rdf
2	schema	rdf	foaf	rdfs
3	gr	foaf	time	owl
4	foaf	dbp	qb	dcterms
5	bibo	dce	scovo	skos
6	time	owl	freebase	foaf
7	skos	rdfs	mo	dce
8	dul	dbo	owl	void
9	ptop	rss	metalex	geo
10	rdafrbr	skos	doap	aktors
11	vaem	gldp	prov	ro
12	ma-ont	geo	void	obo
13	prov	sc	frbr	app
14	swrc	fb	skos	repo
15	dce	gn	dcterms	time

Table 4: Comparing ranking position when using PIC in LOV with respect to prefix.cc and vocab.cc

The LODStats ranking is focused on covering the number of datasets reused in the linked open data cloud [2], which is partially taken into account in our approach. The evidence of that is the first three vocabularies used (`RDF, RDFS, OWL`) which are considered as the meta model for designing vocabularies. Those vocabularies are not included into the LOV catalogue and they do not appear in our ranking. The relative stable position of `foaf` in the four columns of the table suggests that there are equal popular terms. In addition, two other vocabularies have "relative" similar ranking using PIC and LODStats: `skos` and `dcterms`. Regardless the metric used, a short list of the "most popular vocabularies" based on their presence in the Top-15 of the four catalogues is: `foaf, skos` followed by `dcterms, time, dce, prov`.

Closer to our work, Schaible *et al.* reported on an empirical study involving 75 linked data experts and practitioners assessing reuse strategies based on various ranking decisions [6]. The goal is to find objective criteria for choosing which vocabularies to reuse and how many can be combined. LODStats and LOV are used to obtain the number of datasets using a specific vocabulary while *vocab.cc* is used for getting the number of occurrence of a vocabulary term. We propose a different metric to rank existing vocabularies that can be furthermore added as a new feature in such a study. One drawback in the model is to use the same weight for two vocabularies with different number of datasets reused. This could be address in the future by using a "function based" weighting for datasets reused (e.g. inverse logarithm) for computing the PIC score.

[4] `http://vocab.cc/v/tco`

6. CONCLUSION AND PERSPECTIVE

We have presented a different perspective of ranking vocabularies using the principles of Information Content. By applying this concept to Linked Open Vocabularies, we tried to use features that we consider "relevant" to be taken into account when comparing vocabularies (e.g: datasets reused, external vocabularies). We compare with other rankings that are mostly based on the "popularity" of vocabularies. This work can path the way for assessing vocabularies with applications in a more systemic approach for recommending classes/properties in ontology management, or in visualization applications to propose the most *"oh yeah?"* suitable property to be visualized for RDF entities when there is large a large number of properties. As future work, we aim to take into account the equivalence axioms (between classes and properties) when computing the Information Content, and more generally, all sort of semantic relationships between terms. Also, we plan to compare our ranking model with other ranking approaches such as graph-based ones (e.g. pagerank). Another future direction work is to investigate the dependency ranking between vocabularies, by focusing on a specific type of "inlinks" (i.e. extensions, generalization) and study how they affect the PIC values.

Acknowledgment

This work has been partially supported by the French National Research Agency (ANR) within the Datalift Project, under grant number ANR-10-CORD-009.

7. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 2009.

[2] J. Demter, S. Auer, M. Martin, and J. Lehmann. LODStats – An Extensible Framework for High-performance Dataset Analytics. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW'12)*, 2012.

[3] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, and C. V. II. Five stars of linked data vocabulary use. *Semantic Web Journal*, 2014. `http://geog.ucsb.edu/~jano/swj653.pdf`.

[4] R. Meymandpour and J. G. Davis. Ranking Universities Using Linked Open Data. In *5th International Workshop on the Linked Data on the Web (LDOW'13)*, 2013.

[5] S. M. Ross. A First Course in Probability, 2002.

[6] J. Schaible, T. Gottron, and A. Scherp. Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling. In *11th Extended Semantic Web Conference (ESWC'14)*, pages 457–472, 2014.

[7] F. Scharffe and al. Enabling linked-data publication with the Datalift Platform. In *26th AAAI International Conference on Artificial Intelligence (AAAI-12)*, 2012.