An Empirical Study of Vocabulary Relatedness and Its Application to Recommender Systems

Gong Cheng, Saisai Gong, and Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
{gcheng,yzqu}@nju.edu.cn, saisaigong@gmail.com

Abstract. When thousands of vocabularies having been published on the Semantic Web by various authorities, a question arises as to how they are related to each other. Existing work has mainly analyzed their similarity. In this paper, we inspect the more general notion of relatedness, and characterize it from four angles: well-defined semantic relatedness, lexical similarity in contents, closeness in expressivity and distributional relatedness. We present an empirical study of these measures on a large, real data set containing 2,996 vocabularies, and 15 million RDF documents that use them. Then, we propose to apply vocabulary relatedness to the problem of post-selection vocabulary recommendation. We implement such a recommender service as part of a vocabulary search engine, and test its effectiveness against a handcrafted gold standard.

Keywords: Ontology, recommendation, relatedness, vocabulary.

1 Introduction

The Semantic Web enriches data with machine-readable, unambiguous meaning by advising different applications to use common vocabularies (a.k.a. ontologies), and to adhere strictly to the term descriptions provided. It would enable an even wider range of applications that operate on integrated data when vocabularies from different communities are interconnected, e.g. aligned. A large body of work has been devoted to this problem of matching [9], which aims at finding terms (i.e. classes or properties) in different vocabularies that have the same intensional meaning. Accordingly, approaches thus far mainly follow a paradigm that measures the *similarity* between terms [9] or between vocabularies [17,5]. In fact, similarity is just a specific kind of *relatedness*. As other forms of relatedness, one vocabulary may extend another by defining more specific subclasses, and two vocabularies may describe closely related domains so that they are often used together, etc. However, this more general notion of relatedness has been addressed by only few work [19,23,11], and none of these approaches has been evaluated on a representative sample of real-world vocabularies. In this regard, whereas our previous work [4] has analyzed only explicit relations between terms, in this paper, we will characterize several different aspects of relatedness between vocabularies via an empirical study of many real-world, diverse vocabularies.

L. Aroyo et al. (Eds.): ISWC 2011, Part I, LNCS 7031, pp. 98–113, 2011.

Vocabulary relatedness can find many applications. For example, it could be employed to rank and find central vocabularies [7]. Here we conceive another application called *post-selection vocabulary recommendation*. Assume that a user has shown an interest in a vocabulary, or in other words, she has *selected* a vocabulary. Such selection widely exists in many scenarios, e.g. having selected a vocabulary for further exploration when interacting with a vocabulary search engine, or having selected a vocabulary for use when developing an application. Then, a recommender system will automatically suggest several other vocabularies that the user might also be interested in, e.g. one as an alternative or complementary to the selected one for a particular use. Naturally, such recommendation mainly relies on the features of the selected vocabulary, and thus we call it post-selection recommendation. We will discuss how this specific task can be supported by the study of vocabulary relatedness.

To summarize, the contribution of this paper is threefold:

- Rather than similarity, we study the more general notion of relatedness between vocabularies on the Semantic Web. We discuss four kinds of relatedness: (a) semantic relatedness defined by vocabulary (meta-)descriptions, (b) content similarity which exploits lexical features, (c) expressivity closeness according to the language constructs adopted, and (d) distributional relatedness derived from vocabulary usage.
- We apply six proposed relatedness measures to a real-world data set crawled by a Semantic Web search engine, which contains 2,996 vocabularies instantiated by other 15 million RDF documents (collectively containing 4 billion RDF triples). We analyze and compare the effects of our measures, and report many statistical findings that help characterize real-world vocabularies.
- We consider the problem of post-selection vocabulary recommendation, and propose to solve it by using relatedness measures. We also examine the popularity of vocabularies for recommendation. We evaluate our approach based on a handcrafted gold standard, and also develop such a recommender system and incorporate it into a vocabulary search engine.

In the remainder of this paper, Sect. 2 characterizes our data set, in particular the vocabularies identified from it. Section 3 describes and compares several relatedness measures. Section 4 introduces and evaluates a solution to the problem of post-selection vocabulary recommendation. Finally, Sect. 5 compares related work, and Sect. 6 concludes the paper.

2 Vocabularies in the Real World

2.1 Data Set

The data set investigated in this work is the one — at the time of writing — used by the Falcons search engine.[1] As summarized in Table 1, it comprises

[1] `http://ws.nju.edu.cn/falcons/`

15 million RDF (including RDF/XML and RDFa) documents, which collectively contain 4 billion RDF triples, crawled from 5 thousand pay-level domains[2] between February 2010 and May 2011.

Table 1. Data set statistics

Number of RDF documents	15,947,721
Number of pay-level domains hosting RDF documents	5,805
Aggregate number of RDF triples	4,099,414,887
Number of vocabularies	2,996
Number of pay-level domains hosting vocabularies	261
Aggregate number of classes	396,023
Aggregate number of properties	59,868

To characterize the data set, Figure 1 presents the distribution of the number of pay-level domains over the number of RDF documents hosted on a log-log scale. The distribution approximates a power law, but having a long tail to the right which corresponds to several large data sources including `hi5.com`, `l3s.de`, `geonames.org`, `dbpedia.org`, etc. This power law phenomenon has also been observed on other data sets such as the one crawled by Swoogle [8].

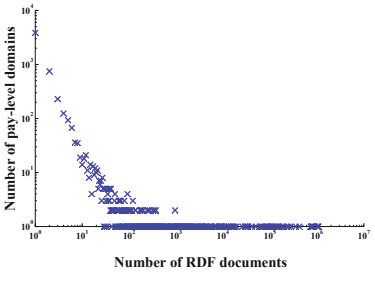

Fig. 1. Distribution of the number of pay-level domains over the number of RDF documents hosted

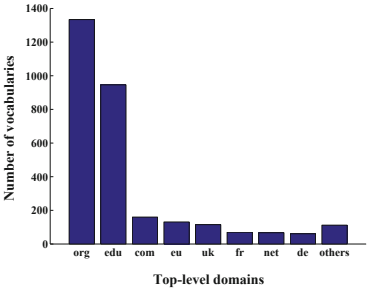

Fig. 2. Distribution of the number of vocabularies hosted over top-level domains, where "others" represents an aggregate of all the ones not presented

2.2 Identifying Vocabularies

We study only the vocabularies that are published by applying best practice.[3] Accordingly, since a vocabulary description may be distributed among multiple

[2] A *pay-level domain* is a domain that requires payment at a (country-code) top-level domain [14]. For instance, the URI `http://ws.nju.edu.cn/falcons/` belongs to the pay-level domain `nju.edu.cn`. We use the Apache Nutch package (`nutch.apache.org`) to identify the pay-level domain of a URI.

[3] `http://www.w3.org/TR/swbp-vocab-pub/`

documents, we employ a bottom-up strategy to identify vocabularies from the data set. That is, firstly we identify a *term* as a dereferenceable URI that refers to a class or a property in the RDF document retrieved via dereferencing the URI. Then, terms sharing a common namespace URI are grouped into a *vocabulary*, using this namespace URI as its identification. In this way, we may miss some old-fashioned vocabularies that are not dereferenceable, and may also fail to find all the terms for some vocabulary, but we believe that the results obtained would accurately reflect real-world conditions at our best.

As summarized in Table 1, we have identified 396,023 classes and 59,868 properties, which are grouped into 2,996 vocabularies. They come from 261 pay-level domains or 33 top-level domains. That is, among 5,805 pay-level domains in our data set that serve RDF documents, only a small portion (4.50%) publish their own vocabularies. Figure 2 depicts the distribution of the number of vocabularies hosted over top-level domains, in which `org` and `edu` dominate with 44.53% and 31.58%, respectively, followed by `com` and several country-code ones. This distribution is also close to the one for Swoogle [8].

These vocabularies vary considerably in size and composition. The largest ones, in terms of the number of terms, are some versions of YAGO and Cyc which comprise tens of thousands of terms, whereas most of the others (72.30%) contain not more than 25. Even among large vocabularies, some (e.g. YAGO) mainly provide classes when some others (e.g. SUMO) are rich in both classes and properties.

3 Characterizing Relatedness between Vocabularies

In this section, we discuss, from different points of view, four kinds of relatedness between vocabularies, and formalize them as numerical measures. In particular, we assume that relatedness measures are symmetric. We perform an empirical analysis of these measures, and finally make a comparison.

3.1 Semantic Relatedness

Vocabularies on the Semantic Web are described in a structured way. When one vocabulary is connected to another via a typed link, it naturally indicates certain kind of relatedness having well-defined semantics, and this leads to our first kind of relatedness measure.

Explicit Relation. Major vocabulary languages such as OWL provide mechanisms for describing information about a vocabulary itself. For instance, `owl:imports` references another vocabulary whose meaning will be included in the present one. Since such relation between vocabularies is directly given in the meta-description of a vocabulary, we call it *explicit relation*. Further, when there are explicit relations between vocabulary v_1 and v_2, and between v_2 and v_3, we observe some kind of relation between v_1 and v_3, which looks "longer" and thus is probably weaker than the two original relations.

These observations could be represented as an edge-weighted graph G_E, where nodes correspond to vocabularies, and every pair of explicitly related vocabularies v_i and v_j are connected by an undirected edge, associated with a weight w indicating how weak the relation is:

$$w(v_i, v_j) = \begin{cases} 2 & \text{if } v_i \text{ references } v_j \text{ or } v_j \text{ references } v_i, \\ 1 & \text{if } v_i \text{ references } v_j \text{ and } v_j \text{ references } v_i. \end{cases} \quad (1)$$

Then, the relatedness (denoted by R_S^E) between two vocabularies is defined as the multiplicative inverse of the weight of a shortest path between their corresponding nodes in G_E, which is thus inside (0,1], or 0 when unreachable. Note that we actually ignore the specific types of relations, as we will see later that most relations observed in practice are quite homogeneous.

Implicit Relation. In a vocabulary, the description of a term may refer to terms in other vocabularies, e.g. via `rdfs:subClassOf` or complex OWL constructs, which suggests a kind of *implicit relation* between vocabularies, in the sense that they are revealed by term-level descriptions but might not be mentioned in the meta-description of vocabulary. Analogous to G_E, here we devise another edge-weighted graph G_I to convey such relations, which differs from G_E in only one respect that: implicit but not explicit relation is considered. Then, a relatedness measure, denoted by R_S^I, is defined based on G_I analogously.

Hybrid Relation. When we take both explicit and implicit relations into consideration, we obtain a kind of *hybrid relation* between vocabularies. Analogously, it could be characterized as an edge-weighted graph G_{E+I}, based on which a relatedness measure, denoted by R_S^{E+I}, is defined.

Empirical Analysis. Among 2,996 vocabularies in the data set, explicit, implicit and hybrid relations are observed between 2,968, 2,845 and 4,691 pairs of vocabularies, respectively. According to Table 2 which summarizes several statistical properties of G_E, G_I and G_{E+I}, whereas G_E and G_I are similar in terms of the number of edges, G_E seems more fragmented, suggested by the percentages of isolated nodes and the metrics below for characterizing reachability. On the other hand, there are far more edges in G_{E+I} than in G_E, indicating that many implicit relations between vocabularies are not captured by the meta-descriptions thereof.

In particular, only 17 types of explicit relations are observed in our data set, and only 6 occur in the meta-descriptions of more than one vocabulary. As shown in Table 3, when `owl:imports` dominates largely, most others are negligible.

3.2 Content Similarity

In a vocabulary description, terms are not only interconnected but also usually associated with human-readable contents, e.g. labels. Given two vocabularies

Table 2. Statistical properties of G_E, G_I and G_{E+I}

	G_E	G_I	G_{E+I}
Number of nodes	2,996	2,996	2,996
Number of edges	2,968	2,845	4,691
Average degree	1.98	1.90	3.13
Maximum degree	786	684	848
Percentage of isolated nodes	56.88%	36.72%	32.31%
Number of connected components	1,763	1,143	1,007
Percentage of nodes in the largest connected component	32.78%	57.44%	62.18%
Percentage of pairs of connected nodes	5.40%	16.50%	19.33%

Table 3. Relations used in the highest percentages of vocabulary meta-descriptions

`http://www.w3.org/2002/07/owl#imports`	36.58%
`http://www.daml.org/2001/03/daml+oil#imports`	1.60%
`http://www.w3.org/2000/01/rdf-schema#seeAlso`	0.30%
`http://www.w3.org/2002/07/owl#priorVersion`	0.10%
`http://purl.org/dc/terms/requires`	0.07%
`http://www.openlinksw.com/schema/attribution#isDescribedUsing`	0.07%

modeling the same or related domains, their textual descriptions often overlap. By detecting this aspect, we present our second kind of relatedness measure.

Specifically, the relatedness (denoted by R_C) between two vocabularies v_i and v_j combines the *content similarity* between their classes (denoted by C_i and C_j) and the one between their properties (denoted by P_i and P_j):

$$R_C(v_i, v_j) = \begin{cases} \frac{\text{SetSim}(C_i,C_j)+\text{SetSim}(P_i,P_j)}{2} & \text{if } C_i \times C_j \neq \emptyset \text{ and } P_i \times P_j \neq \emptyset, \\ \text{SetSim}(C_i, C_j) & \text{if } C_i \times C_j \neq \emptyset \text{ and } P_i \times P_j = \emptyset, \\ \text{SetSim}(P_i, P_j) & \text{if } C_i \times C_j = \emptyset \text{ and } P_i \times P_j \neq \emptyset, \\ 0 & \text{if } C_i \times C_j = \emptyset \text{ and } P_i \times P_j = \emptyset, \end{cases} \tag{2}$$

where SetSim is a similarity measure for term sets that determines the extent to which the lexical features of both sets are covered by each other:

$$\text{SetSim}(T_i, T_j) = \text{HMean}\left(\frac{1}{|T_i|}\sum_{t_i \in T_i} \max_{t_j \in T_j} \text{LS}(t_i, t_j), \frac{1}{|T_j|}\sum_{t_j \in T_j} \max_{t_i \in T_i} \text{LS}\left(t_i, t_j\right)\right), \tag{3}$$

where HMean returns the harmonic mean of the two parameters, and $\text{LS}(t_i, t_j)$ gives the lexical similarity between terms. As one implementation of LS, we apply a string metric [24] to all pairs of the respective labels of the two terms, normalize each result to be inside the interval [0,1], and finally take the maximum.

Empirical Analysis. In our data set, to exploit term descriptions for labels, we retrieve property values from `rdfs:label`, `dc:title` and their subproperties (e.g. `skos:prefLabel`) that are defined via or can be inferred from the `rdfs:subPropertyOf` relation, which collectively amount to 86 types of properties. In this way, at least one label can be found for 63.67% of all the terms, which are distributed among 36.21% of all the vocabularies. Since the absence of label is still commonly observed, the local name of each term URI is also employed.

Another thing we would like to point out is: computing content similarity is the most expensive task performed in our experiments, which costs a multithreading program running on a multi-core server several weeks. This is not surprising because all pairs of 2,996 vocabularies are compared, and for each pair, every class (resp. property) in one vocabulary is compared with every class (resp. property) in another, which is again time-consuming in particular for large vocabularies, as illustrated in Sect. 2.2.

3.3 Expressivity Closeness

Vocabularies vary from lightweight taxonomies to heavyweight ones with complex constraints. In this regard, two vocabularies are close when they are similar in expressivity. Accordingly, we develop our third kind of relatedness between vocabularies based on their *expressivity closeness*.

The expressivity of a vocabulary is mainly (though not fully) captured by the language constructs (e.g. `rdfs:subClassOf` vs. `owl:complementOf`) adopted for describing terms. Besides, other meta-level terms may also be employed for description, e.g. Dublin Core metadata terms and those for meta-modeling. Therefore, we propose to characterize the expressivity of a vocabulary v by MetaTerms(v) — the set of all meta-level terms that are instantiated in v's description. Then, given two vocabularies v_i and v_j, we define their relatedness (denoted by R_E) as follows:

$$R_E(v_i, v_j) = \mathrm{J}(\mathrm{MetaTerms}(v_i), \mathrm{MetaTerms}(v_j))\,, \tag{4}$$

where J returns the Jaccard similarity coefficient of the two sets.

Empirical Analysis. We observe 4,978 meta-level terms that are instantiated in at least one vocabulary's description, 469 (9.42%) of which are used in at least two, showing a wide variety. In particular, the meta-level terms instantiated in the highest percentages of vocabulary descriptions are all language constructs, led by `rdf:type`, `rdfs:domain` and `rdfs:range`. Excluding these, Table 4 presents the top-ranked ones remaining, which are all not widely used.

On the other hand, describing a vocabulary needs to instantiate an average of 10.13 types of meta-level terms. In fact, 92.96% of all the vocabularies in our data set use not more than 20 types. However, we still recognize hundreds of types of meta-level terms in some complex vocabularies such as Cyc.

Table 4. Meta-level terms (excluding those in RDF, RDFS, OWL or DAML) instantiated in the highest percentages of vocabulary descriptions

`http://purl.org/dc/elements/1.1/description`	1.50%
`http://purl.uniprot.org/core/encodedIn`	0.90%
`http://www.w3.org/2004/02/skos/core#definition`	0.73%
`http://purl.org/dc/terms/modified`	0.67%
`http://www.swop-project.eu/ontologies/pmo/product.owl#unit`	0.67%
`http://purl.org/dc/terms/issued`	0.63%
`http://www.w3.org/2003/06/sw-vocab-status/ns#term_status`	0.63%

3.4 Distributional Relatedness

Whereas all the previous notions of relatedness rely on the *intensional* descriptions of vocabularies, our fourth kind of measure looks at the *extensional* side, i.e. to investigate vocabulary usage in practice.

Recall that on the fruitful topic of relatedness in the field of computational linguistics, among others, *distributional relatedness* [20] defines close words as those that are used in similar contexts, e.g. having many co-occurring words in common. Accordingly, a "distributional profile" is created for each word, which characterizes the strength of association between the word and every other word that co-occurs with it, commonly by using conditional probability. Then, the similarity (e.g. cosine) between distributional profiles is calculated, as a proxy for relatedness between words.

Inspired by this line of research, we study vocabulary co-occurrence in use, which in the context of the Semantic Web amounts to vocabulary co-instantiation. We conceive an RDF document as the context from which co-instantiation is observed, and let $\mathrm{IV}(d)$ be the set of all vocabularies instantiated in RDF document d. Then, given the set of all vocabularies V and $v \in V$, the distributional profile of v is represented by a $|V|$-dimensional vector, denoted by $\mathrm{DP}(v)$, where:

$$\mathrm{DP}_i(v) = \frac{|\{d \in D |\, v, v_i \in \mathrm{IV}(d)\}|}{|\{d \in D |\, v \in \mathrm{IV}(d)\}|}, \tag{5}$$

where D is the set of all RDF documents under investigation. In particular, $\mathrm{DP}(v)$ is defined as $\mathbf{0}$ when no instantiation of v can be observed in any $d \in D$. Finally, the relatedness between vocabulary v_i and v_j, denoted by $R_D(v_i, v_j)$, is given by the cosine similarity between $\mathrm{DP}(v_i)$ and $\mathrm{DP}(v_j)$.

This straightforward implementation is improved in two ways. Firstly, language-level vocabularies (e.g. RDF) are trivially and widely instantiated, which function as stop words in computational linguistics. Hence they are filtered out prior to processing. Otherwise, they may undesirably, even largely, increase the relatedness between many pairs of vocabularies. Secondly, as discussed in Sect. 2.1, considering the distribution of the number of pay-level domains over the number of RDF documents hosted, a large data source in the long tail of the

distribution may unfairly affect the computation of relatedness. To avoid this, we limit the effects that could be caused by a single pay-level domain. Specifically, we define $\mathrm{PLD}(D)$ as a partition of D such that each element of $\mathrm{PLD}(D)$ corresponds to all the RDF documents in D that are hosted by one particular pay-level domain. Then, we rewrite (5) as follows:

$$\mathrm{DP}_i(v) = \frac{|\{S \in \mathrm{PLD}(D) | \exists d \in S, v, v_i \in \mathrm{IV}(d)\}|}{|\{S \in \mathrm{PLD}(D) | \exists d \in S, v \in \mathrm{IV}(d)\}|}. \quad (6)$$

Empirical Analysis. In our data set, instantiation is observed for 1,874 (62.55%) vocabularies. Table 5 shows the most widely instantiated ones, led by Dublin Core metadata vocabularies and FOAF. Further, among 9,763 pairs of vocabularies that have co-instantiation, Table 6 presents the most frequent ones.

Table 5. Vocabularies (excluding RDF, RDFS, OWL and DAML) instantiated in RDF documents hosted by the highest percentages of pay-level domains

`http://purl.org/dc/elements/1.1/`	37.45%
`http://xmlns.com/foaf/0.1/`	22.79%
`http://purl.org/dc/terms/`	15.90%
`http://www.icra.org/rdfs/vocabularyv03#`	10.65%
`http://www.w3.org/2003/01/geo/wgs84_pos#`	5.22%
`http://purl.org/vocab/bio/0.1/`	2.76%
`http://www.w3.org/2000/10/swap/pim/contact#`	2.76%
`http://rdfs.org/sioc/ns#`	2.20%
`http://usefulinc.com/ns/doap#`	1.67%
`http://purl.org/vocab/relationship/`	1.38%

Table 6. Pairs of vocabularies (excluding those involving RDF, RDFS, OWL or DAML) co-instantiated in RDF documents hosted by the highest percentages of pay-level domains

`http://purl.org/dc/elements/1.1/` `http://purl.org/dc/terms/`	14.42%
`http://purl.org/dc/elements/1.1/` `http://www.icra.org/rdfs/vocabularyv03#`	10.65%
`http://purl.org/dc/terms/` `http://www.icra.org/rdfs/vocabularyv03#`	10.61%
`http://xmlns.com/foaf/0.1/` `http://purl.org/dc/elements/1.1/`	9.42%
`http://www.w3.org/2003/01/geo/wgs84_pos#` `http://xmlns.com/foaf/0.1/`	5.05%

3.5 Comparison

Now we study the levels of *agreement* between different relatedness measures. We apply, to all pairs of 2,996 vocabularies in our data set, each of our six relatedness measures, namely R_S^E, R_S^I, R_S^{E+I}, R_C, R_E and R_D. Each measure will induce a ranking of these pairs, and we leverage the Spearman's rank correlation coefficient, denoted by ρ, to measure the correspondence between these rankings and assess its significance. ρ is inside the interval [-1,1], and an increasing value implies increasing agreement.

The results are summarized in Fig. 3. All the values are positive, i.e., all these measures are positively correlated. Larger values are found between R_S^I and R_S^{E+I} (0.88), and between R_S^E and R_S^{E+I} (0.53), which are not surprising since R_S^{E+I} comprises R_S^E and R_S^I. In particular, the second largest value (0.66) is observed between R_S^E and R_D, indicating that explicitly related vocabularies are also most likely to be instantiated together, and vice versa.

	R_S^I	R_S^{E+I}	R_C	R_E	R_D
R_S^E	0.39	0.53	0.21	0.19	0.66
R_S^I	-	0.88	0.26	0.38	0.35
R_S^{E+I}	-	-	0.30	0.26	0.43
R_C	-	-	-	0.32	0.23
R_E	-	-	-	-	0.24

Fig. 3. Levels of agreement between individual relatedness measures

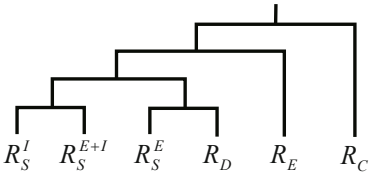

Fig. 4. Dendrogram showing the single-link hierarchical clustering of individual relatedness measures based on their levels of agreement

Further, based on ρ values, we employ the single-link hierarchical clustering technique to depict the relationships between measures. As shown in Fig. 4, R_C is relatively far from the other measures. One reason might be that for describing the same domain, different authorities may publish their own vocabularies, which vary considerably in expressivity and are rarely connected to each other.

4 Post-selection Vocabulary Recommendation

In this section, we describe an application that can be enriched by the study of vocabulary relatedness. Recall that when we browse online book stores or movie databases, some of these applications will provide recommendations to avoid overloading users with information. For instance, when we look at the introduction of a book, several "related items" are also presented, e.g. books written by the same author. Analogously, when interacting with a vocabulary repository, e.g. a vocabulary search engine, after a vocabulary has been selected for examining details, the system is expected to recommend several related vocabularies. In the next, we address this problem of *post-selection vocabulary recommendation*. We describe an approach as well as an extension, and present evaluation results.

4.1 Relatedness-Based Ranking

In Sect. 3, we have introduced six measures of relatedness between vocabularies, namely $\mathfrak{R} = \{R_S^E, R_S^I, R_S^{E+I}, R_C, R_E, R_D\}$, all returning values inside the interval [0,1]. For a selected vocabulary v_0, we argue that a vocabulary v_i is more likely to be recommended if it is more *related* to v_0, in terms of some $R_j \in \mathfrak{R}$. That is, we rank recommendation candidates by $R_j(v_i, v_0)$. Here, which R_j to use is specified by users according to their specific needs. $\mathfrak{R}$ can also be extended to include other metrics developed in the future.

When users intend to receive recommendations featuring several different characteristics, it requires employing multiple measures. Further, users may attach different degrees of importance to different measures. To this end, we allow ranking recommendation candidates by a *linear combination* of all the measures in $\mathfrak{R}$, i.e. $\sum_{R_j \in \mathfrak{R}} \alpha_j R_j(v_i, v_0)$, where $\alpha_j \in [0, 1]$ is a group of weightings.

We implement such a *recommender service* in Falcons Ontology Search.[4] When exploring a retrieved vocabulary, users could enquire about related ones after specifying a weighting for each relatedness measure.

4.2 Popularity-Based Re-ranking

Besides relatedness, another factor we would like to consider in vocabulary recommendation is *popularity*. Recall that the Semantic Web could facilitate data integration on the semantic level exactly because different Semantic Web applications produce and consume data adhering to common vocabularies. Hence, we argue that a recommender service should return more popular vocabularies, i.e. those having been used by more applications. To incorporate popularity into the criteria for ranking, given $\mathrm{Pop}(v)$ — the number of pay-level domains hosting RDF documents that instantiate v, we extend our approach to rank recommendation candidates by the following metric:

$$\sum_{R_j \in \mathfrak{R}} \alpha_j R_j(v_i, v_0) \cdot (1 + \log_b (1 + \mathrm{Pop}(v_i))) , \tag{7}$$

where b is a parameter that tunes the degree of influence of popularity on recommendation. When decreasing b from $+\infty$ to a small value (e.g. 2), the degree of influence increases. But apparently, popularity is achieved at the relative cost of relatedness. A trade-off needs to be studied for specific applications.

4.3 Evaluation

Firstly, without considering popularity, we examine which $R_j \in \mathfrak{R}$ is more useful for recommendation. To achieve this, we compare generated rankings thereof to the gold standard given by human experts. We identify 1,302 vocabularies from our data set for this experiment, each containing 5–25 terms, being neither

[4] `http://ws.nju.edu.cn/falcons/ontologysearch/`

too small to be significant nor too large for manual investigation. We choose 20 from them at random as "selections" for testing post-selection recommendation. For each selection, we can hardly ask experts to give a ranking of all the other 1,301 vocabularies, but rather, we apply the depth-10 pooling technique, which is widely adopted for evaluating information retrieval (IR) systems. To be specific, we apply each $R_j \in \mathfrak{R}$ to score all the other 1,301 vocabularies, retain only those having positive relatedness values, and collect the top-10 ones. For all $R_j \in \mathfrak{R}$, these top-ranked vocabularies collectively form a pool to be used in the experiment. The pool is randomly divided up and assigned to two experts. For each assignment, the expert is asked to assess the relatedness between the assigned vocabulary and the selection, and report (a) "closely related", (b) "somewhat related", or (c) "unrelated", corresponding to ratings 2, 1 and 0, respectively. In particular, 5 vocabularies in each pool are assigned to both experts.

We receive 739 assessments in total, of which 81.60% are unrelated, 10.55% somewhat related and 7.85% closely related. Unrelated vocabularies take the largest proportion, which in fact is quite common under pooling methods. Besides, among 100 (20×5) vocabularies assessed by both experts, agreement is reached on 80%. If we consider only binary ratings by taking closely and somewhat related as "related", agreement is reached on 91%, suggesting a high quality of the assessments. Finally, to form one single gold standard, when two experts give different assessments on a vocabulary, we take the higher rating.

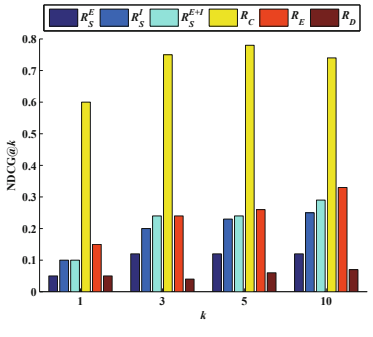

Fig. 5. NDCG of individual measures

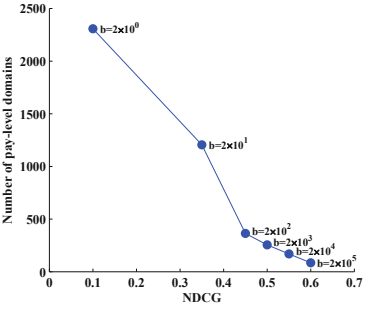

Fig. 6. Relationship between relatedness and popularity under different b values

For each selection, we evaluate each $R_j \in \mathfrak{R}$ by calculating its normalized discounted cumulative gain (NDCG) — a widely used metric for IR evaluation. NDCG@k, inside the interval [0,1], measures the quality of the k top-ranked vocabularies against their gold-standard ratings. Figure 5 summarizes the results averaged over all the 20 selections, under different settings of k. R_C noticeably outperforms the others, showing that our experts assess relatedness between vocabularies mainly based on the overlap between their contents. On the other hand, we attribute the bad performances of R_S^E and R_D to the fact that, as

presented in Sect. 3, 56.88% of vocabularies in our data set are not explicitly related to any other ones, and that 37.45% have no instantiation. Thereby, R_S^E and R_D fail to find any related vocabularies for 13 and 11 selections, respectively. In these cases, NDCG is defined as 0, which largely hurt their overall performances.

Secondly, we look at combinations of measures. Since R_C performs the best in the first experiment, we combine it with every other measure in $\mathfrak{R}$ to see whether better results can be achieved. Figure 7 illustrates the evaluation results of several combinations. Actually, for each kind of combination, we show only one group of weightings with which the best result is obtained. We find that under different settings of k, better or equal results are consistently observed when R_C is combined with R_S^E, R_E or R_D, whereas R_S^I and R_S^{E+I} seem only helpful when $k = 1$, i.e. in generating the top-ranked vocabulary. However, the reader is reminded that these results only reflect the bias of our experts, whereas our flexible approach indeed allows task-oriented combination.

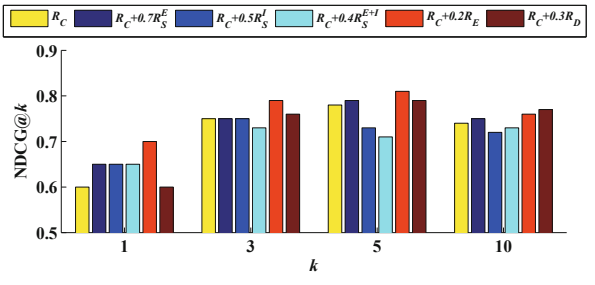

Fig. 7. NDCG of several combinations of measures

Finally, we illustrate, with R_C, the relationship between relatedness and popularity. Under different b values in (7), we evaluate relatedness by NDCG@1 and evaluate popularity by the number of pay-level domains hosting RDF documents that instantiate the top-ranked vocabulary. As shown in Fig. 6, when NDCG decreases from 0.60 to 0.45 (averaged over all the 20 selections), the number of pay-level domains increases linearly. A much higher popularity can also be achieved, which however loses relatedness considerably. It reveals that looking for a good trade-off is not an easy task, but has to rely on specific applications.

5 Related Work

5.1 Relatedness

In computational linguistics, a substantial amount of research has been conducted on the *measurement of relatedness* between words (or senses) [3]. Most existing methods exploit semantic networks such as WordNet, and operate on shortest paths or information theory. These ideas have also been transplanted to the Semantic Web for measuring relatedness between terms in a vocabulary [19,23,11], by treating a vocabulary description as a semantic network.

Complementary to this, another line of research studies the co-occurrence of words to measure their distributional relatedness [20].

Differently, we look at *relatedness on the vocabulary but not the term level*. In an earlier work [4], we derive a vocabulary dependence graph from the relations between terms, and perform a complex network analysis, which reveals its scale-free nature. In [7], several types of relations between terms and between vocabularies are identified, to characterize a random surfer's behavior for ranking. In [25], vocabularies are clustered based on their use of language constructs. Whereas each of these studies investigates very few kinds of relatedness, the work in this paper characterizes it in four aspects and compares six measures.

As a special kind of relatedness, *similarity* between terms [9] and between vocabularies [17,5] have attracted extensive research. Further, the similarities among a collection of vocabularies can be represented as a graph, on which statistical analysis [10,21] and complex network analysis [12] have been carried out. Besides, more sophisticated measures of similarity have been established based on such graph [6]. In our work, we also implement content-based similarity as one aspect, when we deal with the more general notion of relatedness.

5.2 Recommendation

Recommender systems have become an important research area [1]. In particular, collaborative approaches have been applied to vocabulary recommendation [22,15], which are grounded on *user-generated ratings*. A closely related problem is vocabulary search, which usually takes a keyword query as input and in fact performs *query-biased recommendation* [2,13,18]; these approaches mainly investigate how well a vocabulary is relevant to a keyword query. Inspired by [16], the problem of post-selection recommendation addressed in our work is in a different context that takes a selected vocabulary as input and demands *selection-biased recommendations*, to which relatedness measurement is the natural solution.

6 Conclusions and Future Work

In this paper, we have discussed vocabulary-level relatedness from four aspects. Our empirical analysis on a large, real data set compares six developed measures, and also, reports many statistical findings, which help characterize vocabularies on the real Semantic Web. In particular, we observe that many cross-vocabulary relations between terms are not embodied in their vocabulary meta-descriptions, and vocabularies having explicit relations tend to be instantiated together. After that, we have proposed to apply relatedness measures to the problem of post-selection vocabulary recommendation. The evaluation results demonstrate the effectiveness of our measures in recommendation, particularly when they are combined appropriately. We have enriched our Falcons Ontology Search system with such a flexible recommender service.

In fact, our relatedness measures have not fully exploited vocabularies. As future work, textual descriptions and provenance information in vocabulary

meta-description still need investigation. About vocabulary recommendation, it would be interesting to combine our relatedness measures with collaborative methods and ontology evaluation techniques.

Acknowledgments. This work was supported in part by the NSFC under Grant 60973024 and 61021062, and in part by ZTE Corp. (R&Dcon1105160003). We thank Min Liu for his time and effort in the experiments.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6), 734–749 (2005)
2. Alani, H., Brewster, C.: Ontology Ranking Based on the Analysis of Concept Structures. In: 3rd International Conference on Knowledge Capture, pp. 51–58. ACM, New York (2005)
3. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Comput. Linguist. 32(1), 13–47 (2006)
4. Cheng, G., Qu, Y.: Term Dependence on the Semantic Web. In: Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 665–680. Springer, Heidelberg (2008)
5. David, J., Euzenat, J.: Comparison Between Ontology Distances (Preliminary Results). In: Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 245–260. Springer, Heidelberg (2008)
6. David, J., Euzenat, J., Šváb-Zamazal, O.: Ontology Similarity in the Alignment Space. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 129–144. Springer, Heidelberg (2010)
7. Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and Ranking Knowledge on the Semantic Web. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 156–170. Springer, Heidelberg (2005)
8. Ding, L., Finin, T.: Characterizing the Semantic Web on the Web. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 242–257. Springer, Heidelberg (2006)
9. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
10. Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N., Musen, M.A.: What Four Million Mappings Can Tell You about Two Hundred Ontologies. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 229–242. Springer, Heidelberg (2009)
11. Hawalah, A., Fasli, M.: A Graph-based Approach to Measuring Semantic Relatedness in Ontologies. In: International Conference on Web Intelligence, Mining and Semantics, pp. 29:1–29:12. ACM, New York (2011)
12. Hu, W., Chen, J., Zhang, H., Qu, Y.: How Matchable Are Four Thousand Ontologies on the Semantic Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 290–304. Springer, Heidelberg (2011)
13. Jonquet, C., Musen, M.A., Shah, N.H.: Building a Biomedical Ontology Recommender Web Service. J. Biomed. Semant. 1(suppl.1), S1 (2010)

14. Lee, H.-T., Leonard, D., Wang, X., Loguinov, D.: IRLbot: Scaling to 6 Billion Pages and Beyond. In: 17th International Conference on World Wide Web, pp. 427–436. ACM, New York (2008)
15. Lewen, H., d'Aquin, M.: Extending Open Rating Systems for Ontology Ranking and Reuse. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS(LNAI), vol. 6317, pp. 441–450. Springer, Heidelberg (2010)
16. Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., Chang, Y.: Learning to Model Relatedness for News Recommendation. In: 20th International Conference on World Wide Web, pp. 57–66. ACM, New York (2011)
17. Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
18. Martínez Romero, M., Vázquez -Naya, J.M., Munteanu, C.R., Pereira, J., Pazos, A.: An Approach for the Automatic Recommendation of Ontologies Using Collaborative Knowledge. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6277, pp. 74–81. Springer, Heidelberg (2010)
19. Mazuel, L., Sabouret, N.: Semantic Relatedness Measure Using Object Properties in an Ontology. In: Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 681–694. Springer, Heidelberg (2008)
20. Mohammad, S., Hirst, G.: Distributional Measures of Concept-distance: A Task-oriented Evaluation. In: 2006 Conference on Empirical Methods in Natural Language Processing, pp. 35–43. ACL, Sydney (2006)
21. Nikolov, A., Motta, E.: Capturing Emerging Relations between Schema Ontologies on the Web of Data. In: 1st International Workshop on Consuming Linked Data. CEUR Workshop Proceedings (2010)
22. Noy, N.F., Guha, R., Musen, M.A.: User Ratings of Ontologies: Who Will Rate the Raters? In: 2005 AAAI Spring Symposium, pp. 56–63. The AAAI Press, Menlo Park (2005)
23. Pirró, G., Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 615–630. Springer, Heidelberg (2010)
24. Stoilos, G., Stamou, G., Kollias, S.: A String Metric for Ontology Alignment. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 624–637. Springer, Heidelberg (2005)
25. Tempich, C., Volz, R.: Towards a Benchmark for Semantic Web Reasoners-An Analysis of the DAML Ontology Library. In: 2nd International Workshop on Evaluation of Ontology-based Tools. CEUR Workshop Proceedings (2003)