

Technical Report: Property Collocation in DBpedia Ontology: An Empirical Study

Saisai Gong, Gong Cheng, and Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing 210023, PR China
ssgong@smail.nju.edu.cn, {gcheng, yzqu}@nju.edu.cn

Abstract. Some properties of an entity could be grouped or combined to form a facet of the entity description. Grouping or clustering properties from a given ontology will make the ontology more useful to human beings and Web applications. In this paper, we investigate the collocation of properties for grouping properties. We characterize the collocation of properties from three angles: statistical association, semantic collocation and lexical similarity. We present an empirical study of these measures in the DBpedia 3.9 ontology. Results show the existence of property collocation in DBpedia 3.9 ontology, and the property collocation can be characterized better by the linear combination of the three measures.

1 Introduction

Billions of RDF triples describing various entities have been published as Linked Data. In these triples, thousands of properties from different ontologies have been used to describe entities. Some properties of an entity could be grouped or combined to form a facet of the entity description. For example, `longitude` and `latitude` are usually presented together to deliver the geography information of an entity. Grouping properties is very useful in many Linked Data applications. It can be used in ontology search to recommend related properties that users may also need. It also plays an important role in creating navigation structures for entity browsing, which can enhance user experience of browsing [7]. For example, in our Linked Data browser called SView¹ (cf. Section 2), users are free to create their favorite views (i.e., templates) for browsing entities by choosing and arranging properties, and share these views between each other. Automatically offering a good initial grouping of properties can save users' time. Therefore, it is necessary to group or cluster properties from a given ontology, and it will make the ontology more useful to human beings and Linked Data applications.

There are a large number of ontologies of different sizes available on the Web. Some ontologies contain a thousand or even more properties. For example, the DBpedia 3.9 ontology has 1,650 properties [11]. In this case, automatically grouping properties is usually adopted for scalability rather than manual grouping. To automatically group properties in ontologies, the critical step is to determine the

¹ <http://ws.nju.edu.cn/sview2/>

relatedness among the properties. There have been many researches investigating the measures of property relatedness. Most of them focus on property similarity in order to efficiently identify *synonymous* or *equivalent* properties [1,20]. However, the properties grouped together are not only the synonymous or equivalent ones, but also the ones sharing similar topics in most cases, such as longitude and latitude. In addition, synonymous or equivalent properties are just prevalent across different ontologies due to the heterogeneity of the ontologies, but it is infrequent that they appear in the same ontology.

In this paper, we study *property collocation* for grouping properties. Property collocation refers to a combination of properties that happens very often and more frequently than expected. With collocation between properties, properties can be grouped using clustering methods. We will show that the property collocation in ontologies is common and can be measured. We characterize the property collocation from three angles: statistical association, semantic collocation and lexical similarity. We present an empirical study of these measures in the DBpedia 3.9 ontology. Results show the existence of property collocation in DBpedia 3.9 ontology, and the property collocation can be characterized better by the linear combination of the three measures.

The rest of this paper is structured as follows. Section 2 gives an application scenario for property collocation. Section 3 gives an overview of our approach. Section 4 introduces different measures of property collocation. Our evaluation is reported in Section 5. Related work is discussed in Section 6. We conclude this paper in Section 7.

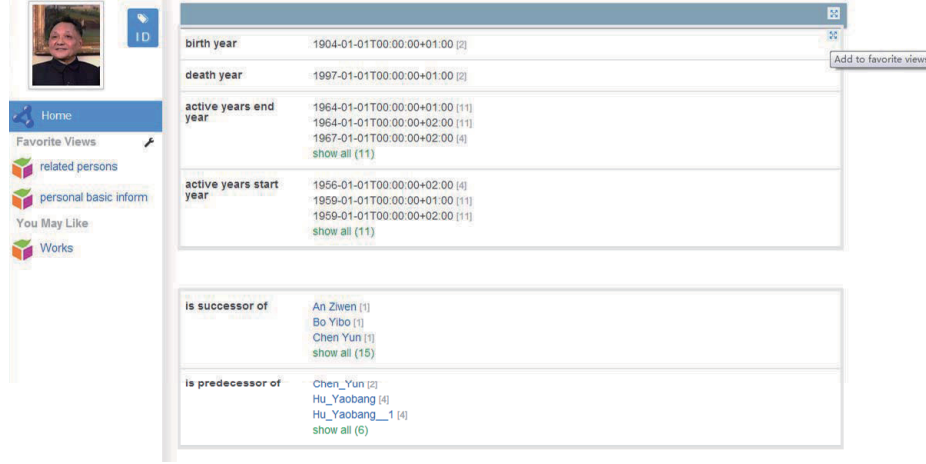
2 An Application Scenario for Property Collocation

In this section, we present an application of property collocation in entity browsing. In our Linked Data browser named SView, when a user browses an entity, in order to make entity descriptions more meaningful, she can create her favorite views of the entity descriptions freely by choosing and arranging entity properties, as shown in Fig. 1. Each view consists of several entity properties representing a facet of the entity descriptions. The views will be shared by users when needed. In this way, users can also bookmark recommended views from other users or collect part of the content in them to form their favorite views.

Automatically offering a good initial grouping of properties can save users' time and alleviate user burden. To achieve it, SView generates initial property groups by clustering the properties based on property collocation. With these initial property groups, users can adjust these groups by adding or removing properties and then collect the groups to form their favorite views.

For example, in the Fig. 1, when a user browses the entity named Deng Xiaoping, SView automatically generates two property groups {**birth year**, **death year**, **active years end year**, **active years start year**} and {**is successor of**, **is predecessor of**}. The user can add these two groups to an existing view or create a new one with these two groups. She can also split

the first group into two groups {active years end year, active years start year}, {birth year, death year} and collect them to form a favorite view.




 ID	
birth year	1904-01-01T00:00:00+01:00 [2]
death year	1997-01-01T00:00:00+01:00 [2]
active years end year	1964-01-01T00:00:00+01:00 [11] 1964-01-01T00:00:00+02:00 [11] 1967-01-01T00:00:00+02:00 [4] show all (11)
active years start year	1956-01-01T00:00:00+02:00 [4] 1959-01-01T00:00:00+01:00 [11] 1959-01-01T00:00:00+02:00 [11] show all (11)
is successor of	An Zhen [1] Bo Yibo [1] Chen Yun [1] show all (15)
is predecessor of	Chen_Yun [2] Hu_Yaobang [4] Hu_Yaobang__1 [4] show all (6)

Fig. 1. User interface of the SView

3 Overview of Our Approach

In this paper, we focus on the collocation among properties in the same ontology. As for the collocation of properties from different ontologies, due to the terminological heterogeneity in different ontologies, we believe that the identification of property collocation within an ontology proposed in this paper should be combined with *ontology matching* (property matching) techniques to identify property collocation across ontologies, which will be our future work.

To investigate collocation among the properties from a given ontology, we firstly identify collocational property pairs, and then cluster these property pairs to form collocational property groups if necessary. The key issue is to introduce good quantitative measures of the collocation between two properties for effectively identifying collocational pairs. We characterize the collocation between properties from the following three angles: statistical association, semantic collocation and lexical similarity.

Statistical association. In the field of computational linguistics, statistical association [6] is used to identify collocation between words based on word co-occurrence in different contexts, such as bigrams, sentences or Web documents. Based on the selected context, the strength of association or collocation between co-occurring words is quantified by using mutual information or other measures. Inspired by this line of research, we study property co-occurrence in usage, i.e.,

co-instantiation in the context. We conceive an entity’s description as the context from which co-instantiation is observed, i.e., both properties are used to describe the entity. Then, the statistical association in this context will be measured.

Semantic collocation. Property axioms in ontologies may indicate strength of collocation between properties to some extent. For example, collocated properties usually have similar *domain* and/or *range*. In other words, properties with similar domain or range are more likely to be collocated. In view of this, our semantic collocation between properties is developed based on the property axioms in ontologies, if there are.

Lexical similarity. Properties are usually associated with human-readable textual descriptions, e.g., labels. It is natural that the similarity of textual descriptions may also indicate some kind of collocation. Considering this aspect, we leverage several string metrics and WordNet-based measures to quantify the lexical similarity of properties’ textual descriptions and use it as a proxy for the collocation between properties.

In the evaluation, we also combine the three types of measures by using a linear combination for comparison. In the next section, we will describe these three types of measures in detail.

4 Characterizing Collocation between Properties

4.1 Statistical Association

We use three popular statistical association measures to characterize property collocation: Phi Coefficient (denoted by C_ϕ), Symmetrical Uncertainty Coefficient (denoted by C_U) and Jaccard Coefficient (denoted by C_J).

Phi Coefficient is a measure of the degree of association between two binary variables, which is a special case of the Pearson Correlation Coefficient. Its value is limited to the interval $[-1, +1]$; a higher positive value indicates a stronger association. Let $P(p_i)$ be the probability that an entity is described by the property p_i (Recall that we use an entity as the context of the property co-occurrence, as described in Section 3), which could be estimated by counting throughout a corpus. Let $P(\bar{p}_i) = 1 - P(p_i)$ be the probability that an entity is not described by the property p_i . Analogously, let $P(p_i, p_j)$ be the probability that an entity is described by both the properties p_i and p_j , $P(p_i, \bar{p}_j)$ be the probability that an entity is described by p_i but not by p_j , $P(\bar{p}_i, \bar{p}_j)$ be the probability that an entity is described neither by p_j nor by p_j . Phi Coefficient of p_i and p_j is computed by the following formula:

$$C_\phi(p_i, p_j) = \frac{P(p_i, p_j)P(\bar{p}_i, \bar{p}_j) - P(\bar{p}_i, p_j)P(p_i, \bar{p}_j)}{\sqrt{P(p_i)P(p_j)P(\bar{p}_i)P(\bar{p}_j)}} \quad (1)$$

Symmetrical Uncertainty Coefficient is also a measure of the strength of association between variables, which is based on information theory. It is a

normalized mutual information. Its value is in the range $[0,1]$; a higher value indicates a stronger association. Let $H(p_i) = -\sum_{x \in \{p_i, \bar{p}_i\}} P(x) \log P(x)$ be the entropy of the property p_i and $H(p_i, p_j) = -\sum_{x \in \{p_i, \bar{p}_i\}} \sum_{y \in \{p_j, \bar{p}_j\}} P(x, y) \log P(x, y)$ be the joint entropy of p_i and p_j . Symmetrical Uncertainty Coefficient of p_i and p_j is computed as follows:

$$C_U(p_i, p_j) = 2 \times \frac{H(p_i) + H(p_j) - H(p_i, p_j)}{H(p_i) + H(p_j)} \quad (2)$$

Jaccard Coefficient measures the similarity between two sets. Its value ranges from 0 to 1; a higher value indicates a stronger association between properties in our scenario. Jaccard Coefficient of p_i and p_j is computed by the following formula:

$$C_J(p_i, p_j) = \frac{P(p_i, p_j)}{P(p_i) + P(p_j) - P(p_i, p_j)} \quad (3)$$

Note that to calculate the values of the above measures, we need to obtain the probabilities $P(p_i)$, $P(p_i, p_j)$, $P(p_i, \bar{p}_j)$ and $P(\bar{p}_i, \bar{p}_j)$. Let U be a set of URI references, B be a set of blank nodes and L be a set of literals. A triple $(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$ is called an RDF triple. An RDF graph is a set of RDF triples. Let the RDF graph G be the corpus from which the probabilities are estimated. Let $\text{Res}(G)$ be the entities in the graph G :

$$\text{Res}(G) = \{s \mid \exists p, o, (s, p, o) \in G\} \cup \{o \in U \cup B \mid \exists s, p, (s, p, o) \in G\} \quad (4)$$

Let $\text{ResDescr}(G, p_i)$ be the entities that are described by the property p_i :

$$\text{ResDescr}(G, p_i) = \{s \mid \exists o, (s, p_i, o) \in G\} \cup \{o \in U \cup B \mid \exists s, (s, p_i, o) \in G\} \quad (5)$$

The probabilities used by the above measures are estimated as follows:

$$\begin{aligned} P(p_i) &= \frac{|\text{ResDescr}(G, p_i)|}{|\text{Res}(G)|}, \\ P(p_i, p_j) &= \frac{|\text{ResDescr}(G, p_i) \cap \text{ResDescr}(G, p_j)|}{|\text{Res}(G)|}, \\ P(p_i, \bar{p}_j) &= \frac{|\text{ResDescr}(G, p_i) \cap (\text{Res}(G) - \text{ResDescr}(G, p_j))|}{|\text{Res}(G)|}, \\ P(\bar{p}_i, \bar{p}_j) &= \frac{|\text{Res}(G) - (\text{ResDescr}(G, p_i) \cup \text{ResDescr}(G, p_j))|}{|\text{Res}(G)|}. \end{aligned} \quad (6)$$

4.2 Semantic Collocation

We leverage the domain, range and subproperty axioms in the ontology to characterize semantic collocation between properties, if there are.

Assume that the domain or range of a property is `rdfs:Resource` if there is no explicit domain or range axiom of the property in the ontology. Let $d(p_i)$ ($r(p_i)$) be the domain (range respectively) of the property p_i (Note that $d(p_i)$ and

$r(p_i)$ are two class sets), and $dmin(p_i)$ ($rmin(p_i)$ respectively) be the minimal classes in $d(p_i)$ ($r(p_i)$ respectively), i.e. $dmin(p_i) = \{c \in d(p_i) \mid \nexists x \in d(p_i), x \text{ is a direct or indirect subclass of } c\}$. Given two properties p_i, p_j defined in the ontology \mathcal{O} , the semantic collocation of p_i and p_j w.r.t the property axioms in \mathcal{O} (denoted by C_S) is calculated as follows, the value of which is in the range $[0,1]$:

$$C_S(p_i, p_j) = \alpha SetSim_d(dmin(p_i), dmin(p_j)) + \beta HRel(p_i, p_j) + \gamma SetSim_r(rmin(p_i), rmin(p_j)) \quad (7)$$

where $SetSim_d$ and $SetSim_r$ denote the similarity of two minimal class sets for property domain and range respectively, $HRel$ is a relatedness measure based on the property hierarchy defined in \mathcal{O} , and $\alpha, \beta, \gamma \in [0,1], \alpha + \beta + \gamma = 1$. We will detail $SetSim_d$, $SetSim_r$ and $HRel$ in the following.

Let $len_p(\cdot, \cdot)$ denote the length of a shortest path between two properties in the property hierarchy, and let $mss(\cdot, \cdot)$ denote the set of the most specific super properties of two certain properties. $HRel(p_i, p_j)$ is computed as follows:

$$HRel(p_i, p_j) = \frac{1}{1 + \min_{x \in mss(p_i, p_j)} \max(len_p(x, p_i), len_p(x, p_j))} \quad (8)$$

The value of $HRel(p_i, p_j)$ is higher when p_i is a direct subproperty of p_j , or p_i and p_j are sibling properties. Note that if p_i and p_j have no common super properties, $HRel(p_i, p_j) = 0$.

Let $len_c(x, y)$ denote the length of a shortest path between two classes x, y in the class hierarchy defined in the global ontology consisting of \mathcal{O} and all \mathcal{O}' s (directed and indirected) imported ontologies. Let S_i, S_j denote two class sets. We define a bipartite graph $G_{S_i, S_j} = (V, E)$ such that $V = V_{S_i} \cup V_{S_j}$, $|V| = |S_i| + |S_j|$, each vertex $v \in V_{S_i}$ ($v \in V_{S_j}$) corresponds to a class $c \in S_i$ ($c \in S_j$ respectively), and for each edge $e = (x, y) \in E$, $x \in V_{S_i}, y \in V_{S_j}$, and the edge weight $w(e) = \frac{1}{len_c(x, y) + 1}$. We denote $mwm(S_i, S_j)$ as the sum of the weights of the edges in the maximum weighted matching of G_{S_i, S_j} . The similarity of two minimal class sets of property domain $SetSim_d$ in Eq. (7) is computed in three different ways (denoted by $SetSim_d^J$, $SetSim_d^{max}$, $SetSim_d^{mwm}$):

$$SetSim_d^J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \quad (9)$$

$$SetSim_d^{max}(S_i, S_j) = \max_{x \in S_i, y \in S_j} \frac{1}{len_c(x, y) + 1}. \quad (10)$$

$$SetSim_d^{mwm}(S_i, S_j) = \frac{2 \times mwm(S_i, S_j)}{|S_i| + |S_j|} \quad (11)$$

Except $SetSim_d^J(S_i, S_j)$, both $SetSim_d^{max}(S_i, S_j)$ and $SetSim_d^{mwm}(S_i, S_j)$ consider the subclass relationships of the classes in S_i and S_j . $SetSim_d^{max}(S_i, S_j)$ only uses the closest class pair in S_i and S_j while $SetSim_d^{mwm}(S_i, S_j)$ considers all the class pairs.

Let $\mathbf{S} = \{\text{rdfs:Resource}\}$. The similarity of two minimal class sets of property range $SetSim_r$ in Eq. (7) is defined as follows:

$$SetSim_r(S_i, S_j) = \begin{cases} 1 & S_i = S_j = \mathbf{S} \\ \frac{2 \times mwm(S_i, S_j)}{|S_i| + |S_j|} & S_i \neq \mathbf{S} \text{ and } S_j \neq \mathbf{S} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

The value of $SetSim_r(S_i, S_j)$ is 1 when both p_i and p_j have no explicit range axioms in the ontology ($S_i = S_j = \mathbf{S}$). If the ranges of p_i and p_j are both non-trivial, i.e. having a subclass of `rdfs:Resource`, $SetSim_r(S_i, S_j)$ is computed using $\frac{2 \times mwm(S_i, S_j)}{|S_i| + |S_j|}$.

4.3 Lexical Similarity

To characterize lexical similarity between properties for collocation identification, we take advantage of three string similarity measures: I-Sub similarity (denoted by C_I), JaroWinkler similarity (denoted by C_R), Levenshtein similarity (denoted by C_L), and a WordNet-based string similarity measure (denoted by C_W). The values of C_R and C_L are in range $[0,1]$ while the value of C_I is in $[-1,+1]$; a larger value indicates a higher similarity. The textual descriptions of properties used in this paper for computing lexical similarity are the property labels.

I-Sub is a string metric that measures string similarity by considering both commonalities and differences of two strings [17]. The detailed calculation of the I-Sub can be found in [17].

JaroWinkler is also a measure of similarity between two strings. It is designed and best suited for short strings such as person names. The detail calculation of JaroWinkler similarity can be found in [18].

Levenshtein similarity is derived from the Levenshtein distance. The Levenshtein distance is a string metric for measuring the difference between two strings, which is the minimum number of edit operations required to transform one string to another. Levenshtein similarity of two strings s_i and s_j is computed as follows:

$$C_L(s_i, s_j) = 1 - \frac{LevenshteinDistance(s_i, s_j)}{\max(|s_i|, |s_j|)} \quad (13)$$

WordNet can be treated as a knowledge base for measuring lexical semantic relatedness between words. Various WordNet-based similarity measures between words have been proposed [2], based on shortest paths, information theory, etc. To calculate the WordNet-based string similarity of two properties p_i, p_j , we first transform each property label into a normalized form by removing stop words, stemming and so on. Let l_i and l_j be the resulting normalized form of p_i and p_j respectively, and let $|l_i|$ and $|l_j|$ are the number of words in l_i and l_j respectively. $C_W(p_i, p_j)$ is calculated by the following formula:

$$C_W(p_i, p_j) = \min\left(\frac{\sum_{x \in l_i} \max_{y \in l_j} WuP(x, y)}{|l_i|}, \frac{\sum_{x \in l_j} \max_{y \in l_i} WuP(x, y)}{|l_j|}\right) \quad (14)$$

where $WuP(\cdot, \cdot)$ is the Wu and Palmer’s WordNet-based lexical similarity between words defined in [19], which operates on the shortest path.

5 Evaluation

In the previous sections, we have introduced several measures for property collocation. In this section, we will firstly present a study of human identification of property collocation in the DBpedia 3.9 ontology ², which investigated whether property collocation exists in the DBpedia ontology and how common it is. Then, we will report the comparison results of the proposed collocation measures based on the results of the human identification.

The screenshot shows a web interface for a survey. At the top, there are tabs for 'Visited: 70 / 70', 'Visited entities', and 'Questionnaire'. Below these are buttons for 'Save Grouping' and 'Next Entity'. The main content area is titled 'Montana State University' and contains a description of the university. Below the description, there are two columns of properties. The left column, 'Ungrouped properties', lists various properties of the university. The right column, 'Grouped properties', lists properties that are grouped together. A tooltip is visible over the 'is alma mater of' property in the grouped list, indicating that it can be moved to a group.

Ungrouped properties		Grouped properties	
abstract	Montana State University (MSU) is a public university located in Bozeman, Montana, United States. It is the state's land-grant university and primary campus in the Montana State University System, which is part of the Montana University System. MSU o... More	alma mater	
affiliation	Big Sky Conference National Collegiate Athletic Association	is education of	Brittany Wiser Jake Whittenberg Susan Roesgen
campus	College town	is college of	Al Wilson (Canada) Harvey Wylie Johnathan Taylor (show all (8))
endowment	1.0025423E8	is alma mater of	Albert Spaulding Arlene Becker Bill Thomas (Monta show all (32))
mascoat	Bobcats : Champ	number of students	
motto	Mountains & Minds	number of postgraduate students	1924
official school colour	Blue and Gold	number of undergraduate students	10840
president	Waded Cruzado		
type	Land-grant university		

Fig. 2. User interface of the survey

5.1 Human Identification of Collocation

Before comparing measures of property collocation, we must firstly answer the following questions: Whether property collocation in the ontology exists? How common it is? How people agree on property collocation? In order to better understand the problem of collocation identification and answer the questions, we performed a study of human identification of property collocation in the DBpedia 3.9 ontology, since the properties in the DBpedia ontology cover different domains and are more apt to form different property groups.

² Available at <http://wiki.dbpedia.org/Downloads39>

Survey setup. In order to perform the study, we delivered an online survey ³ to acquire human judgments of the property collocation in the DBpedia 3.9 ontology. The survey invited Web users to browse different entities in the DBpedia, and group the properties of these entities to form the collocational property groups (See Section 3) based on users' own thoughts. Users can put a property into a group by moving or copying, as shown in Fig. 2. In this way, user judgments of the property collocation were obtained. To make diversity of the properties to be grouped, the survey randomly selected 70 entities for browsing from 14 popular classes in the DBpedia (5 entities in each class) such that each entity has at least 20 properties in the <http://dbpedia.org/ontology> namespace. The 14 classes are as follows: Artist, Athlete, Band, Book, Building, Company, EducationInstitution, Film, MusicalWork, Politician, PopulatedPlace, River, Software, and TelevisionShow. When assigning next entity for browsing, the survey always gave priority to the entities that were browsed most infrequently. Every participant of the survey was invited to browse at least 5 entities. The raw data of the survey is available at <http://ws.nju.edu.cn/sview/gpindex.jsp>.

The selected 70 entities for browsing cover 22.6% of the DBpedia ontology's properties (373 in 1,650) in all. We distinguish the properties in *forward* and *backward* direction, and consider the different directions of the same property representing different properties generally. In this regard, the selected 70 entities have 464 properties with direction. The survey collected a dataset of collocational property groups identified by 33 different users in the end. The dataset consists of 508 property groups in total, and contains 98% (455 in 464) of the selected 70 entities' properties in these groups, of which 63.15% (293 in 464) are forward properties. In this dataset, a user identifies 23.7 groups in average, while the maximum number is 124. Some property groups were identified by multiple users. In average, a group is identified by 1.54 users and the maximum is 10.

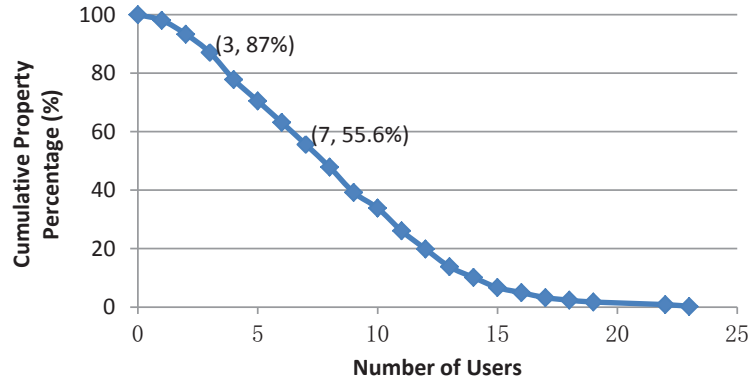


Fig. 3. Cumulative percentage of properties grouped by various number of users

³ <http://ws.nju.edu.cn/sview/gpindex.jsp>

Result analysis. The cumulative percentage of properties grouped by various number of users is shown in Fig. 3. For example, the point (3, 87%) means that 87% (404 in 464) properties were grouped with other ones by at least 3 users since the users believed these properties had collocative partners, i.e. they were collocated with the other properties. From Fig. 3, we can see more than a half (55.6%) properties were regarded as having collocative partners by at least 7 users. This indicates that property collocation is common in the DBpedia ontology.

The cumulative number of groups identified by various number of users is shown in Fig. 4 on a log-linear scale. From this figure, we can see more than 10% groups (62 in 508) were identified by at least 3 users, of which 28 groups were identified by at least 5 users. It shows some degree of agreement on property collocation between users.

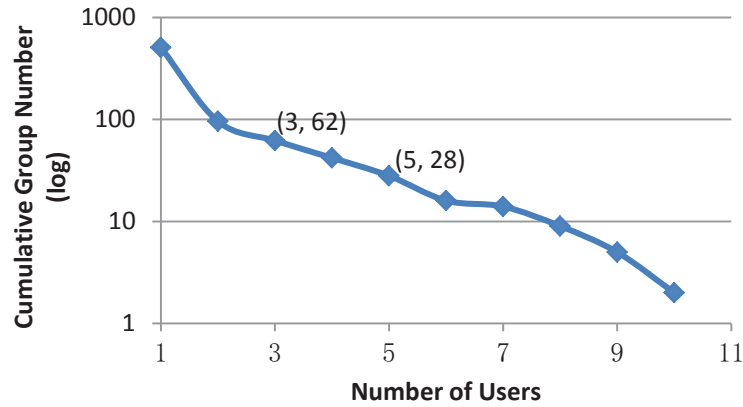


Fig. 4. Cumulative number of groups identified by various number of users

Since the property groups identified by the users were overlapped, to better understand the human identification of property collocation, we also investigated the dataset from the view of the collocational property pairs, which were derived from the property groups. We obtained 12,682 property pairs (with direction) in total regarded as being collocated by at least one user. The cumulative number of property pairs regarded as being collocated by various number of users is shown in Fig. 5 on a log-linear scale. From Fig. 5, we can see about 15% pairs (1,845 in 12,682) were considered as collocated properties by at least 3 users. 576 (about 5% in 12,682) pairs were referred to as being collocated by at least 5 users, which already contain more than a half of all the properties with direction (269 in 464). The pairs identified by at least 3 users contain even more properties, i.e. 77% (358 in 464). All these indicate that more than a half properties in our sample were collocated with some others, and users made agreement on at least one collocative partner of these properties.

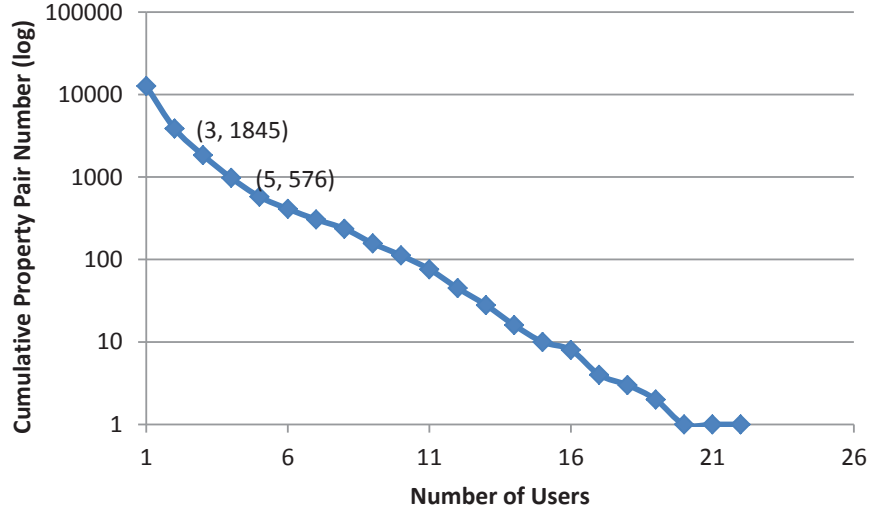


Fig. 5. Cumulative number of collocated pairs identified by various number of users

5.2 Comparison of Different Measures

Dataset. The target of this experiment was to evaluate how the property collocation in the ontology could be characterized by the statistical association, semantic collocation and lexical similarity. We leverage the dataset of the property collocation from human identification as described in Section 5.1 to compare the different measures proposed in Section 4. To compute statistical association, we used a corpus consisting of the datasets called “Mapping-based Properties”, “External links”, “Redirects”, “Disambiguation links”, “Page IDs”, “Revision IDs” in the DBpedia 3.9. The reason why we chose these datasets was that the instantiation information of all the 70 entities’ properties was contained in them and we focused on the collocation between the properties in the <http://dbpedia.org/ontology> namespace.

Experiment setup. We compared different measures in terms of the Spearman’s rank correlation coefficient between their ratings of the collocational property pairs (See Section 5.1) and the human rating. We first calculated human ratings of the collocational property pairs identified by different number of users, ranging from 3 to 12, i.e. the pairs identified by at least 3, 4, \dots , 12 users. For each number, we sorted the property pairs identified by at least that number of users in descending order of the values of the sorting function $Sort(\cdot, \cdot)$, while the ties were broken by the alphabetical order of the property URIs and the directions. Given a property pair (p_i, p_j) , let $N(p_i, p_j)$ be the set of users identifying the pair (p_i, p_j) , and $M(p_i)$ be the set of users grouping the property p_i , i.e. the set of users identifying at least one property collocated with p_i . The function $Sort$

is calculated as follows:

$$Sort(p_i, p_j) = \frac{|N(p_i, p_j)|}{|M(p_i) \cup M(p_j)|} \quad (15)$$

In this way, we obtained 10 lists of the property pairs representing the human rating. Let L_i ($3 \leq i \leq 12$) denote the list of the property pairs identified by at least i users. For each L_i , we calculated the different measures' ratings of the pairs in L_i by ranking the pairs in descending order of the measure values and breaking ties via property URIs and directions, and compared them with the human rating. Table 1 shows the correlation coefficients between the human rating and different measures' ratings for the L_i ($3 \leq i \leq 12$), where C_S^{mwm} , C_S^{max} , C_S^J represents the semantic collocation measure C_S computed using $SetSim_d^{mwm}$, $SetSim_d^{max}$ and $SetSim_d^J$ respectively, as described in Section 4.2.

Table 1. Correlation coefficients between human rating and different measures

List	C_U	C_J	C_ϕ	C_S^{mwm}	C_S^{max}	C_S^J	C_I	C_L	C_R	C_W
L_3	0.280	0.265	0.272	0.246	0.246	0.268	0.202	0.331	0.327	0.154
L_4	0.286	0.277	0.277	0.295	0.295	0.292	0.380	0.392	0.416	0.143
L_5	0.426	0.400	0.416	0.250	0.250	0.227	0.457	0.435	0.453	0.205
L_6	0.519	0.490	0.500	0.245	0.245	0.224	0.466	0.420	0.456	0.200
L_7	0.563	0.533	0.532	0.212	0.212	0.207	0.447	0.428	0.428	0.221
L_8	0.589	0.559	0.553	0.210	0.210	0.207	0.430	0.446	0.441	0.237
L_9	0.514	0.454	0.463	0.123	0.123	0.121	0.444	0.494	0.476	0.285
L_{10}	0.492	0.405	0.443	-0.003	-0.003	-0.010	0.449	0.520	0.498	0.309
L_{11}	0.436	0.344	0.396	0.017	0.017	0.007	0.343	0.429	0.363	0.268
L_{12}	0.418	0.349	0.386	0.111	0.111	0.097	0.469	0.624	0.484	0.257

Result analysis. From Table 1, we can see the correlation coefficients of statistical association-based (C_U , C_J , C_ϕ) and lexical similarity-based measures (C_I , C_L , C_R) are generally comparable, and both of them are higher than the coefficients of semantic collocation-based measures (C_S^{mwm} , C_S^{max} , C_S^J). This means statistical association or lexical similarity matches human judgments of collocation better when there exists proper corpus for property co-occurrence statistics or meaningful property labels. In the statistical association-based measures, the Symmetrical Uncertainty Coefficient (C_U) reflects human judgments of collocation better with the highest coefficient values. In the semantic collocation-based measures, C_S^{mwm} usually has higher coefficient values. Among the lexical similarity-based measures, the I-Sub (C_I) similarity reflects human judgments best in some cases while the Levenshtein (C_L) similarity generally agrees better with the human judgments in other cases. In addition, the coefficient value of the WordNet-based similarity (C_W) performs worst in our context.

Further, we argue that when the statistical association, semantic collocation and lexical similarity identifying property collocation from different angles are combined, property collocation may be characterized better. To validate it, we selected three measures C_U , C_S^{mwm} and C_L , and investigated the effects of the linear combination of the three measures in terms of the correlation coefficients with the human judgments, i.e. $\alpha_1 C_U + \alpha_2 C_S^{mwm} + \alpha_3 C_L$, where $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$, $\alpha_1 + \alpha_2 + \alpha_3 = 1$, as well as those of the linear combinations of any two of the three measures. For each combination, the parameters α_i were automatically set for obtaining high coefficient value. Table 2 presents the results of different combinations ($0.7C_U + 0.3C_S^{mwm}$, $0.5C_U + 0.5C_L$, $0.7C_L + 0.3C_S^{mwm}$, $0.5C_U + 0.1C_S^{mwm} + 0.4C_L$).

Table 2. Correlation coefficients of linear combinations

List	C_U	C_S^{mwm}	C_L	$C_U + C_S^{mwm}$	$C_U + C_L$	$C_L + C_S^{mwm}$	$C_U + C_S^{mwm} + C_L$
L_3	0.280	0.246	0.331	0.313	0.368	0.376	0.391
L_4	0.286	0.295	0.392	0.350	0.414	0.436	0.436
L_5	0.426	0.250	0.435	0.446	0.516	0.466	0.528
L_6	0.519	0.245	0.420	0.522	0.541	0.454	0.561
L_7	0.563	0.212	0.428	0.570	0.588	0.460	0.611
L_8	0.589	0.210	0.446	0.592	0.611	0.459	0.628
L_9	0.514	0.123	0.494	0.515	0.613	0.493	0.624
L_{10}	0.492	-0.003	0.520	0.446	0.599	0.514	0.605
L_{11}	0.436	0.017	0.429	0.421	0.523	0.420	0.526
L_{12}	0.418	0.111	0.624	0.424	0.633	0.620	0.636

From Table 2, we can see the linear combination of all the three measures achieves the highest coefficient value compared to the others, which indicates the property collocation is characterized better by the combination of the statistical association, semantic collocation and the lexical similarity. In addition, the linear combinations of any two of the three measures generally have higher coefficient values than the single measures, which indicates the statistical association, semantic collocation or the lexical similarity can not be replaced with each other.

6 Related Work

The need to determine the relatedness between two properties or relations is an important problem that attracts many researches in the Semantic Web area. Measures of property relatedness can be used in such applications as ontology matching [16], entity resolution [10], query expansion [1,20] and exploratory search [12].

Property similarity is a special kind of property relatedness. There are a lot of existing works dedicated to the property similarity to find synonymous properties or equivalent properties [1,16,20]. [1] leveraged association rule mining to discovery synonymous properties. The work in [20] presented the statistical knowledge patterns which identified synonymous properties in and across datasets based on the triple overlap, cardinality ratio and clustering. The methods in ontology matching can be used to align properties in general. In our work, we investigate the property collocation which is also a special kind of property relatedness but different from the property similarity: collocated properties are not only the synonymous or equivalent ones, but also the ones sharing similar topics or frequently combined in most cases.

There are also works focusing on the more general notion of relatedness. [8] used the Web as knowledge source and exploited the frequencies of use provided by search engines to define semantic relatedness measure among ontology terms. In [3], the relatedness between vocabularies was characterized from four angles: well-defined semantic relatedness, lexical similarity in contents, closeness in expressivity and distributional relatedness. Two properties that are related may not be collocated ones. For example, two properties are related just because they are both inverse functional properties. However, they are never combined in usage.

One important notion related to the property collocation is word collocation. Word collocation refers to characteristic and frequently recurrent word combinations [6]. It is one of the important issues of linguistics, which is widely used in a variety of natural language processing tasks such as word disambiguation. Inspired by this, we investigate in this paper whether property collocation, i.e. a combination of properties that happens very often and more frequently than expected, exists in the ontologies and how common it is. In computational linguistics, statistical association among words [14] is usually used to identify word collocation. Some of these measures can be adjusted to identify property collocation. Besides using statistical association measures, we also leverage property axioms and lexical similarity to identify property collocation.

Both faceted categorization and clustering organize the items into meaningful groups in order to make sense of the items and help decide what to do next [9]. They are provided by many Linked Data browsing systems to help users explore Linked Data [4]. Automatic construction of facets draws attentions in many studies [13,15]. Faceted categorization and clustering are generally used to group entities while our work focuses on grouping entity properties.

7 Conclusions and Future Work

There are many application scenarios for grouping properties, which make ontologies more useful to human beings and Linked data applications. In this paper, we investigated the property collocation for grouping properties. The main contributions of this paper are as follows:

- We propose property collocation and show the existence and the prevalence of property collocation in the DBpedia ontology by a survey.
- We empirically evaluate three kinds of property collocation measures: statistical association, semantic collocation and lexical similarity. We show property collocation can be characterized using these measures by our experiments.
- We integrate property collocation into an Linked Data browser called SView for grouping properties and forming views.

In the future work, we will investigate property collocation across ontologies by using property collocation and ontology matching. Furthermore, we will cope with the problem of how to improve property grouping by leveraging user feedback.

Acknowledgements This work is supported in part by the NSFC under Grant 61170068, 61223003, and 61100040, and in part by the JSNSF under Grant BK2012723. We would like to thank all of the participants in the survey.

References

1. Abedjan, Z., Naumann, F.: Synonym analysis for predicate expansion. In: *ESWC*, pp. 140–154, 2013
2. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1): 13–47, 2006
3. Cheng, G., Gong, S., Qu, Y.: An empirical study of vocabulary relatedness and its application to recommender systems. In: *ISWC*, pp. 98–113, 2011
4. Dadzie, A.S., Rowe, M.: Approaches to visualising linked data: A survey. *Semantic Web*, 2(2): 89–124, 2011
5. Defays, D.: An efficient algorithm for a complete link method. *The Computer Journal*, 20(4): 364–366, 1977
6. Evert, S.: Corpora and collocations. *Corpus Linguistics: An International Handbook*, 2008
7. Fagan, J.C.: Usability studies of faceted browsing: A literature review. *Information Technology and Libraries*, 29(2):58–66, 2013
8. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: *WISE*, pp. 136–150, 2008
9. Hearst, M.A.: Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4): 59–61, 2006
10. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web*, 23: 2–15, 2013
11. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2013
12. Marchionini, G.: Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4): 41–46, 2006
13. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: *ISWC*, pp. 559–572, 2006

14. Pecina, P., Schlesinger, P.: Combining association measures for collocation extraction. In: *COLING*, pp. 651–658, 2006
15. Sah, M., Wade, V.: Personalized concept-based search and exploration on the web of data using results categorization. In: *ESWC*, pp. 532–547, 2013.
16. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1): 158–176, 2013
17. Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment. In: *ISWC*, pp. 624–637, 2005
18. Winkler, W.E.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In: *SRMS*, pp. 354–359, 1990
19. Wu, Z., Palmer, M.: Verbs semantics an lexical selection. In: *ACL*, pp. 133–138, 1994
20. Zhang, Z., Gentile, A.L., Blomqvist, E., Augenstein, I., Ciravegna, F.: Statistical knowledge patterns: Identifying synonymous relations in large linked datasets. In: *ISWC*, pp. 703–719, 2013