An Empirical Study on Property Clustering in Linked Data[⋆]

Saisai Gong, Haoxuan Li, Wei Hu, and Yuzhong Qu

State Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
ssgong.nju@gmail.com, hxli.nju@gmail.com,
whu@nju.edu.cn, yzqu@nju.edu.cn

Abstract. Properties are used to describe entities, a part of which are likely to be clustered together to constitute an aspect. However, existing automated approaches to property clustering remain far from satisfactory for an open domain like Linked Data. In this paper, we firstly investigated the relatedness between properties using five different measures. Then, we employed three clustering algorithms and two combination methods for property clustering. We empirically studied the property clustering on a moderate-sized sample of Linked Data and found that a proper combination of different measures gave rise to the best result.

1 Introduction

With the development of Linked Data, billions of RDF triples have been published to describe numerous entities. An entity usually involves multiple aspects and its property-values may focus on different aspects. For instance, graduate from and work at reveal the career information of a person, while parent, spouse and child deliver her family information. Therefore, it is natural to cluster properties into meaningful groups based on the aspects that they intend to describe. Property clustering is useful for many applications such as entity browsing, ontology editing, query completion, etc. It makes the presented information more formatted and understandable and significantly enhances the capability of users to consume the large-scale Linked Data. However, automated property clustering for an open domain like Linked Data remains far from satisfactory due to the multi-sourced and heterogeneous vocabularies used.

In this paper, we empirically studied the property clustering in Linked Data. We tried our best in this study to provide answers to the following questions:

Q1. What is the most effective measure(s) for measuring property relatedness?
Q2. What is the most effective algorithm(s) for clustering properties ?
Q3. Can the combination method(s) improve the property clustering and how largely?
Q4. Are there any general principles or guidelines for using the property clustering in practice?

[⋆] Funded by the National Natural Science Foundation of China (No. 61370019).

2 Property Relatedness Measures

To achieve property clustering, we measure the relatedness between properties from the following five perspectives.

- *Lexical similarity between property names*, denoted by R_I, is based on the common characters of property names. For example, both mouth position and mouth elevation describe the mouth information of a river. We calculated R_I using the I-Sub string similarity [3].
- *Semantic relatedness between property names*, denoted by R_W, leverages WordNet to measure the semantic relatedness between properties. We used the average Lin's WordNet relatedness [2] of word pairs in property names to calculate R_W.
- *Distributional relatedness between properties*, referred to as R_U, is based on the property co-occurrence in the context of an entity's RDF description, i.e. both properties are used together to describe the entity. Symmetrical uncertainty coefficient was used to compute the distributional relatedness. To estimate the probabilities of co-occurrence, the Billion Triples Challenge (BTC) 2011 dataset[1] was used, in which the descriptions of coreferent URIs were merged.
- *Range relatedness between properties*, referred to as R_T, is based on the class relatedness of property ranges. For example, if two properties have the ranges delicious food and handicraft respectively, both of them deliver the tourist information of a tourist city. The range relatedness is calculated using the maximum WordNet-based relatedness R_W of class pairs in property ranges.
- *Overlap of property values*, denoted by R_O, leverages the common values of two properties to compute the relatedness. The text of each property value is firstly collected, e.g. local names of URIs and lexical forms of literals after normalization, and all the terms in the text are used to construct a term frequency vector. R_O is then computed using the cosine similarity of the corresponding vectors.

3 Clustering Algorithms and Combination Methods

We employed the following three well-known clustering algorithms: DBSCAN (denoted by C_D), Single linkage clustering (C_L) and Spectral clustering (C_S). Combining various relatedness measures helps obtain a better clustering. We employed two typical combination methods. The first one is to first compute property relatedness using a *linear combination* of different measures for each property pair and then carry out clustering. The second one is to first conduct clustering based on individual measures and then aggregate these individual results using *ensemble clustering*. We selected *consensus clustering* to realize ensemble clustering and calculated it using CC-Pivot [1].

[1] `http://km.aifb.kit.edu/projects/btc-2011/`

Table 1. Average performance w.r.t. relatedness measures and clustering algorithms

(a) Precision

	C_D	C_L	C_S
R_I	.235	.235	.184
R_W	.215	.215	.198
R_U	.242	.242	.177
R_T	.170	.170	.215
R_O	.247	.247	.188

(b) Recall

	C_D	C_L	C_S
R_I	.273	.273	.449
R_W	.266	.266	.337
R_U	.433	.433	.410
R_T	.381	.381	.329
R_O	.137	.138	.427

(c) F-Score

	C_D	C_L	C_S
R_I	.253	.253	.261
R_W	.238	.238	.250
R_U	.310	.310	.248
R_T	.235	.235	.260
R_O	.176	.177	.261

(d) Rand Index

	C_D	C_L	C_S
R_I	.549	.549	.500
R_W	.672	.672	.584
R_U	.644	.644	.503
R_T	.547	.547	.628
R_O	.709	.708	.516

(e) NMI

	C_D	C_L	C_S
R_I	.387	.387	.229
R_W	.441	.441	.231
R_U	.507	.507	.224
R_T	.364	.364	.255
R_O	.520	.520	.216

4 Empirical Study

We report our study of the relatedness measures, clustering algorithms and combination methods. Their clustering performance w.r.t. the golden standard was evaluated using the following five metrics: *Precision*, *Recall*, *F-Score*, *Rand Index* and *Normalized Mutual Information* (NMI). All the parameters were set as the ones achieving the highest harmonic mean of F-Score.

We sampled 20 entities of different types in Linked Data, each of which was integrated from a DBpedia URI with its coreferent ones from 12 different sources[2]. Every entity has at least 51 properties while the maximum number is 574. The golden standard was built based on Freebase. Freebase divides properties describing similar aspects into *types* and groups similar types into *domains*. We invited three PhD candidates in the field of Linked Data to assign each property to the most relevant /domain/type. The properties that were assigned to the same /domain/type were clustered together to form the golden standard. The Fleiss' κ inter-rater agreement score is 0.895, showing the strong agreement.

Table 1 depicts the average performance achieved w.r.t. different measures using clustering algorithms. Overall, no measure achieves the highest values for every clustering algorithm on all the measures. R_I and R_U generally generate better clusterings in terms of F-Score. Besides, from the third column of each table, we saw that C_D is similar to C_L and C_S is greatly different from them. C_D and C_L usually generate better clustering results in terms of Rand Index and NMI. Table 2 shows the harmonic means of Precision, Recall, F-Score, Rand Index and NMI achieved by using single measures, linear measure combinations

[2] These sources are DBpedia, DBTune, Freebase, GeoNames, LinkedGeoData, LinkedMDB, New York Times, OpenCyc, Project Gutenberg, RDF Book Mashup, The World Factbook and YAGO

Table 2. Comparison on single relatedness measures and two combination methods

Clustering algorithm: C_D	Precision	Recall	F-Score	Rand Index	NMI
R_I	.235	.273	.253	.549	.387
R_W	.215	.266	.238	.672	.441
R_U	.242	.433	.310	.644	.507
R_T	.170	.381	.235	.547	.364
R_O	.247	.137	.176	.709	.520
$.3R_I + .7R_U$	.218	.757	.339	.471	.379
$.5R_I + .5R_O$	.209	.619	.313	.411	.265
$.6R_U + .4R_O$	.214	.716	.330	.477	.375
$.3R_I + .5R_U + .2R_O$	.211	.883	.341	.398	.318
$.3R_I + .5R_U + .1R_T + .1R_O$	.205	.878	.333	.372	.277
$.2R_I + .1R_W + .2R_U + .5R_O$	.216	.790	.339	.438	.344
$.2R_I + .1R_W + .15R_U + .1R_T + .45R_O$	.207	.899	.337	.364	.268
R_I, R_U	.287	.148	.196	.732	.563
R_I, R_O	.331	.051	.089	.744	.566
R_U, R_O	.290	.066	.108	.755	.575
R_I, R_U, R_O	.273	.210	.237	.706	.513
R_I, R_U, R_T, R_O	.292	.102	.151	.744	.560
R_I, R_W, R_U, R_O	.290	.115	.165	.726	.548
R_I, R_W, R_U, R_T, R_O	.256	.213	.232	.677	.493

and ensemble clustering (the 13th to 19th rows). The results indicate that the linear combination of relatedness measures tends to generate a clustering that features a higher Recall compared to single measures, while ensemble clustering is recommended to use if a higher Precision is preferred.

5 Conclusion

In this paper, we studied the property clustering in Linked Data and evaluated different property relatedness measures, clustering algorithms and combination methods. Our experimental results demonstrated the feasibility of the automated property clustering. In future work, we will improve the quality of property clustering by leveraging user feedback and active learning.

References

1. Ailon, N., Charikar, M., Newman, A.: Aggregating Inconsistent Information: Ranking and Clustering. Journal of the ACM, 55(5):23 (2008)
2. Lin, D.: An Information-Theoretic Definition of Similarity. In: ICML 1998. pp. 296–304. Morgan Kaufmann, San Francisco (1998)
3. Stoilos, G., Stamou, G., Kollias, S.: A String Metric for Ontology Alignment. In: Gil, Y., et al. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 624–637. Springer, Heidelberg (2005)