CD at ENSEC 2016: Generating Characteristic and Diverse Entity Summaries

Danyun Xu, Liang Zheng, and Yuzhong Qu

National Key Laboratory for Novel Software Technology, Nanjing University, China
{dyxu,zhengliang}@smail.nju.edu.cn,yzqu@nju.edu.cn

Abstract. We introduce our entity summarization approach called CD, which aims to select characteristic and diverse features into an entity summary. The characterizing ability of a feature is measured according to information theory. The information overlap between features considers ontological semantics of classes and properties, as well as string and numerical similarity. Finally, selecting characteristic and diverse features is formulated as a binary quadratic knapsack problem to solve.

Keywords: Entity summarization, self-information, reasoning, string similarity, numerical similarity.

1 Introduction

Our entity summarization approach, called CD, is adapted from [3]. The basic idea is to, given an entity description composed of a set of property-value pairs called features, select a size-limited subset of characteristic and diverse features as an entity summary. We formulate it as a binary quadratic knapsack problem (QKP) to solve. Specifically, the characterizing ability of a feature is measured according to information theory, and the information overlap between features considers ontological semantics of classes and properties, as well as string and numerical similarity.

2 Preliminaries

Let E, C, P, and L be the sets of all entities, classes, properties, and literals in a dataset, respectively. The description of an entity e is a set of property-value pairs called features, denoted by $d(e) \subseteq P \times (E \cup C \cup L)$. In RDF data, $d(e)$ is obtained from RDF triples in which e is the subject or the object. When e is the subject of a triple t, the predicate (which is a property) and the object (which is an entity, a class, or a literal) of t comprise a feature. When e is the object of a triple t, the inverse of the predicate and the subject of t comprise a feature. The inverse of a property p is a property automatically created by our approach and is distinguished from p, though they share a common name; if a property p_i is a subproperty of a property p_j, we also define the inverse of p_i as a subproperty of the inverse of p_j. Given an integer k, an entity summary S of e is a subset of $d(e)$ subject to $|S| \leq k$.

3 Approach

3.1 Characterizing Ability of a Feature

The characterizing ability of a feature f, denoted by $ch(f)$, is measured according to information theory. Specifically, we compute the normalized amount of self-information contained in the probabilistic event of observing f in an entity description in a dataset. A feature will have high characterizing ability if it belongs to a small number of entity descriptions:

$$ch(f) = \frac{-\log \frac{|\{e \in E : f \in d(e)\}|}{|E|}}{\log |E|}\,, \tag{1}$$

which is in the range of $[0, 1]$.

3.2 Information Overlap between Features

The information overlap between two features f_i and f_j, denoted by $ovlp(f_i, f_j)$, considers ontological semantics of classes and properties, as well as string and numerical similarity.

For a feature f, let $prop(f)$ and $val(f)$ return the property and the value of f, respectively.

Firstly, we exploit ontological semantics of classes and properties. If both $prop(f_i)$ and $prop(f_j)$ are `rdf:type` and $val(f_i)$ is a subclass of $val(f_j)$ (or vice versa), we will define $ovlp(f_i, f_j) = 1$ because one of them can be inferred from the other and thus they share maximized overlapping information. Similarly, we will also define $ovlp(f_i, f_j) = 1$ if $val(f_i) = val(f_j)$ and $prop(f_i)$ is a subproperty of $prop(f_j)$ (or vice versa).

In other cases, we calculate the string similarity between property names ($isub$) and the similarity between property values (sim):

$$ovlp(f_i, f_j) = \max\{isub(prop(f_i), prop(f_j)), sim(val(f_i), val(f_j)), 0\}\,, \tag{2}$$

which is in the range of $[0, 1]$. Here, $isub \in [-1, 1]$ returns the ISub string similarity [2] between two property names; $sim \in [-1, 1]$ returns the similarity between two property values. To measure $sim(val(f_i), val(f_j))$, if both $val(f_i)$ and $val(f_j)$ are numerical data values, we calculate their similarity as follows.

1. If $val(f_i) = val(f_j)$, $sim(val(f_i), val(f_j)) = 1$;
2. otherwise, if $val(f_i) \cdot val(f_j) \leq 0$, $sim(val(f_i), val(f_j)) = -1$;
3. otherwise, $sim(val(f_i), val(f_j)) = \frac{\min\{|val(f_i)|, |val(f_j)|\}}{\max\{|val(f_i)|, |val(f_j)|\}}$.

In other cases, we treat $val(f_i)$ and $val(f_j)$ as strings; that is, for entities and classes, we take their names, and for literals, we take their string forms. Then we calculate their ISub string similarity as sim.

3.3 Selecting Characteristic and Diverse Features

We aim to select up to k features from $d(e)$ that maximize their total characterizing ability and minimize the total information overlap between them. To this end, we define the quality of an entity summary S as

$$q(S) = \gamma \cdot \sum_{f \in d(e)} ch(f) + \delta \cdot \sum_{f_i, f_j \in S} -ovlp(f_i, f_j)\,, \tag{3}$$

in which $\gamma, \delta > 0$ are the weights of the two objectives to tune, to achieve different trade-offs.

Maximizing q can be reformulated as an instance of QKP [1] as follows. We number the features in $d(e)$ from f_1 to $f_{|d(e)|}$. By introducing a series of binary variables x_i for $i = 1 \cdots |d(e)|$ to indicate whether f_i is selected into the optimal summary, the problem is formulated as

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{|d(e)|} \sum_{j=i}^{|d(e)|} p_{ij} x_i x_j \\ \text{subject to} \quad & \sum_{i=1}^{|d(e)|} x_i \leq k \\ & x_i \in \{0, 1\} \text{ for } i = 1 \cdots |d(e)|\,, \end{aligned} \tag{4}$$

in which p_{ij} is the "profit" achieved if both f_i and f_j are selected:

$$p_{ij} = \begin{cases} \gamma \cdot ch(f_i) & \text{if } i = j\,, \\ \delta \cdot (-ovlp(f_i, f_j)) & \text{otherwise}\,. \end{cases} \tag{5}$$

We solve QKP using a state-of-the-art heuristic algorithm [4].

Acknowledgments. This work was supported in part by the NSFC under Grant 61572247 and in part by the Fundamental Research Funds for the Central Universities.

References

1. Pisinger, D.: The Quadratic Knapsack Problem - A Survey. Discrete Appl. Math. 155(5), 623–648 (2007)
2. 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, Berlin Heidelberg (2005)
3. Xu, D., Cheng, G., Qu, Y.: Facilitating Human Intervention in Coreference Resolution with Comparative Entity Summaries. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 535–549. Springer International Publishing, Switzerland (2014)
4. Yang, Z., Wang, G., Chu, F.: An Effective GRASP and Tabu Search for the 0-1 Quadratic Knapsack Problem. Comput. Oper. Res. 40(5), 1176–1185 (2013)