Collective Alignment of Large-scale Ontologies

Varun R Embar
UC Santa Cruz
vembar@ucsc.edu

Jay Pujara
University of Southern California
jay@cs.umd.edu

Lise Getoor
UC Santa Cruz
getoor@soe.ucsc.edu

The rapid growth in digitization of data has led to creation of fragmented but vital knowledge sources. Ontologies are one such crucial source of knowledge and aligning them is a key challenge for creating an Open Knowledge Network. The task of ontology alignment has received significant attention. In this abstract, we building on existing work, and propose a novel probabilistic ontology alignment approach that combines several similarity measures with structural information such as subsumption and mutual exclusion.

Most large-scale ontologies such as product catalogs [Agrawal and Srikant 2001] and folksonomies [Plangprasopchok et. al. 2010] do not have a formally defined ontology with well-defined classes, instances and properties. Instead, they loosely define relationships such as subsumption between various entities. For example, a folksonomy for Instagram would contain not only tags corresponding to people, places and activities but also tags such as `Selfie`, which correspond to a type of image. Product catalogs have very different textual representation for the same entity. For instance, products related to 3D printing are present in a category called `3D Printing & Supplies` on Ebay, while the same products are present in a category called `Additive Manufacturing Products` on Amazon. Moreover, the same textual representation might have different semantics based on the source of the ontology. The category `Headphones` in an ontology corresponding to a particular company (such as Bose) is different from the `Headphones` category of a large e-commerce retailer such as Amazon. Even aligning tracks in a music catalog is considerably challenging as it is unclear whether the tracks `Bohemian Rhapsody OST` and `Bohemian Rhapsody Remastered 2011` are the same. To sum up, ontology alignment is challenging due to informally defined subsumptions, multiple textual representations for the same class, ambiguity of similar textual representations and presence of large number of instance variations .

Existing ontology alignment approached can be classified into schema-based approaches, instance-based approaches and hybrid approaches [Euzenat et. al. 2007]. Hybrid approaches such as Information-Flow-based Map [Kalfoglou and Schorlemmer, 2003] combines string-based heuristics and the structure of the ontology to generate alignments. Naive Ontology Mapping [Ehrig and Sure, 2004] makes uses of rules that exploit information present in the ontology.

Motivated by these hybrid methods, our proposed ontology alignment approach combines several similarity and distance scores with soft structural constraints. We then define a probability distribution over the set of all possible alignments that takes into account correlations between different alignments. Apart from similarity scores computed on the textual representation of entities, we also compute scores using the entity hierarchy described by the subsumption relations. This helps in identifying the semantics on each entity. Apart from structural constraints such as mutual exclusion, we also incorporate relation specific constraints. For instance, it is unlikely that multiple entities that have a parent-child relationship align to a single entity. We use Probabilistic Soft Logic(PSL)[Bach et. al. 2017], a powerful probabilistic programming framework, that uses weight first-order logic rules to define a probability distribution. Having defined the distribution, we use the efficient MAP inference supported by PSL to identify the most likely alignment.

We performed experiments on product taxonomies extracted from four websites and compared our method to a tf-idf similarity score based approach. While the instance-based similarity score prevented aligning categories such as `bicycle stands & storage` and `storage & home organization`, the structural constraints helped distinguish between equivalence and more general relations. For example, `beauty & personal care` was aligned to `beauty` and not `hair care`, even though there is a significant overlap of products, as `hair care` was the child of `beauty` in the product taxonomy. In summary, combining multiple scores and structural constraints using a probabilistic framework led to a 36% improvement in precision and a 15% improvement in F1 score over the string similarity baseline.