Unionability

Das Sarma et al. consider two factors in measuring the unionability of domains: entity consistency and entity expansion [1]. In their work, entity consistency is quantified by comparing domains in ontology space. Domains are mapped to classes in an ontology using the semantic relatedness of their entity values to Classes.

In ontology O, a domain $D_i = \{v_i^1, v_i^2, \ldots\}$ is represented as $O(D_i) = \{c_1 : w_i^1, c_2 : w_i^2, \ldots, c_k : w_i^k\}$, where $c_j \in Classes(O)$ and $w_j = f(D_i, c_j)$ is a calculated score, using some function f, for the relatedness of domain D_i to class c_j. The ontology similarity (OS) of a query domain D_q and a candidate domain D_c, is defined as follows:

$$SO(D_q, D_c) = Sim(O(D_q), O(D_c)) \tag{1}$$

Dot-product is the similarity function used by Das Sarma et al.

$O(D_i)$ represents domain D_i in the space of ontology classes. However, c_i's are not orthogonal dimensions of the ontology.

In embedding space, we represent domain $D_i = \{v_i^1, v_i^2, \ldots\}$ as $E(D_i) = \{emb(v_i^1), emb(v_i^2), \ldots\}$, where $emb(v_i)$ is the embedding of the domain value v_i. We define the topic representation of domain D_i as $P(E(D_i)) = \{P_i^1 : v_i^1, \ldots, P_i^k : v_i^k\}$, where P_j is the j-th principal component of $E(D_i)$ and v_j is the variance of P_j. In the embedding topic space, P_j's and v_j's are analogous to c_k's and w_k's in ontology space. However, P_j's are data-dependent and variable for different domains, as opposed to c_j's that are fixed for all domains.

Assume $SE(E(D_q), E(D_c))$ is the semantic similarity of domains D_c and D_q, in topic embedding space. Various similarity measures can be considered for comparing domains in topic embedding space. If Probabilistic PCA is used for generating principal components, log-likelihood of D_c in the topic embedding space of D_q can be used as a semantic similarity measure.

Another way is to find the alignment a between P_i's in $P(E(D_c))$ and P_j's in $P(E(D_q))$ such that $\sum Cosine(a)$ is maximized.

We show that $SE(D_q, D_c)$ is monotonically increasing with respect to $Cosine(nearest(P(E(D_c)), P(E(D_q))))$.

The notion of entity expansion can be quantified by the containment score of query domain and candidate domains.

References

[1] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding related tables. *SIGMOD*, pages 817–828, 2012.