Figure 1: Top Pay-Level Domains (PLDs) of entities in BTC-2009

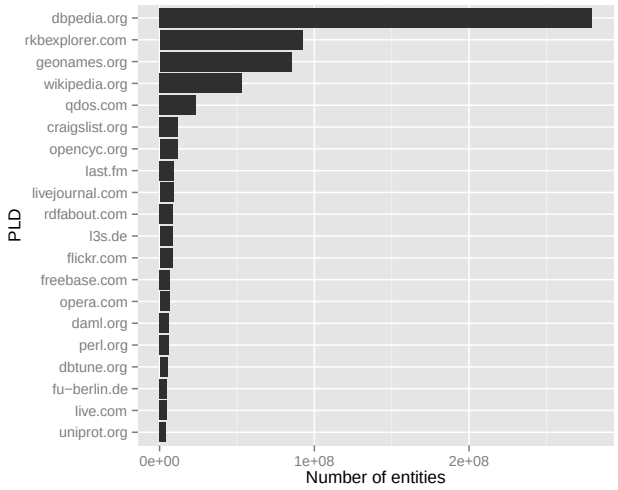
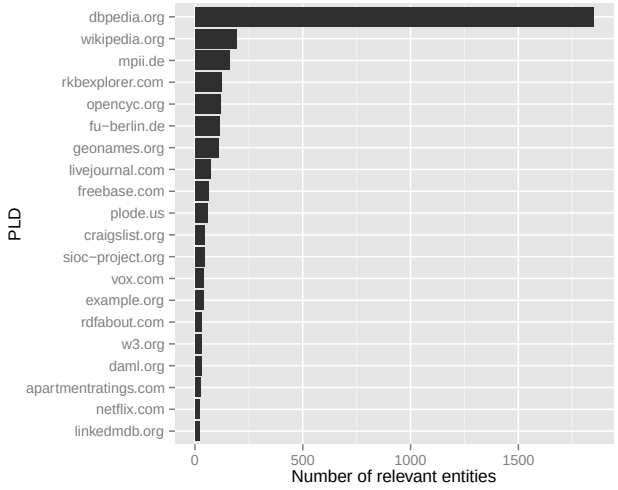


Figure 2: Top Pay-Level Domains (PLDs) of relevant entities in Semantic Search Challenge 2010/2011



1 Top PLDs of entity sources

Figures 1 and 2 show numbers of entities for Top-20 Pay-Level Domains for the whole BTC-2009 dataset and relevant results from SemSearch Challenge judgments only respectively. It can be observed that in BTC-2009 dataset entities are significantly skewed towards DBpedia, and for SemSearch Challenge

this disproportion is even higher as was noted in [1]. Speaking of numbers, in the whole BTC-2009 dataset DBpedia entities constitute 34.6% of all entities and in relevance judgments they constitute 48.9% of all relevant results.

2 Feature usefulness analysis

We've analyzed significance of our features for different types of concepts using one-sided Mann-Whitney test (significance level = 0.01). We've observed that for Field Probability feature for concepts of type *attribute* values of feature for *attributes* field are significantly higher than values for all three names fields (*names*, *similar entity names*, and *related entity names*); for *entity* and *relation* concept types feature value for all names fields is significantly higher than values for both *attributes* and *categories* fields; for *type* concepts Field Probability values for *categories* field is significantly higher than values for all other fields. For Top Score feature values for *attribute* concepts for *attributes* field is higher than values for all other fields; for *relation* concepts values for *similar entity names* are significantly higher than values for all other fields.

References

[1] Krisztian Balog and Robert Neumayer. A test collection for entity search in dbpedia. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 737–740. ACM, 2013.