Leveraging the Emergent Semantics in Twitter Lists for Ontology Development

Andrés García-Silva, Oscar Corcho

Facultad de Informática, Universidad Politécnica de Madrid
{hgarcia,ocorcho}@fi.upm.es

Abstract. In this poster we present an approach to obtain automatically domain ontologies relying on domain vocabularies elicited from Twitter lists and on the reuse of conceptualizations in existing knowledge bases. We tap into relations established between list names, under which prominent users in the domain of study as well as in the microblogging platform have been classified, to harvest related concepts. The relations between concepts are identified from interlinked general-purpose knowledge bases.

1 Introduction

Twitter, the microblogging platform, enables users to organize others into lists. Other users can benefit of existing lists by subscribing to them so that they can receive updates of the people classified under these lists. Given the size of the social network which nowadays reaches 100 million active users, and the bottom-up classification structure emerging from the connection between list names, curators, subscribers, and members, these lists potentially constitutes a valuable resource for knowledge acquisition. In table 1 we presents the terms found in list names under which a journalist have been listed. Note that most of the terms are semantically related around the news domain. In fact, in [1] we have shown that list names are semantically related according to co-occurrence patterns which have been defined in terms of the use given by curators, members and subscribers.

Table 1. Terms found in list names with the corresponding frequency of appearance.

news	297	politics	208	media	58	new_politics	36	celebrities	34
celebs	28	political	26	journalists	25	twibes	18	national	17

Therefore, we aim at collecting a vocabulary, relevant in the domain of study, from the classification structure emerging from Twitter lists. We reuse conceptualizations in existing knowledge bases so that we can define the semantics of the relations between the terms in the vocabulary. With the lists of terms and relations we create an ontology schema which we may also populate with instances from the knowledge bases.

2 Approach

2.1 Preprocessing

During this activity we extract, normalize, and transform the lists. For data *extraction* we rely on the REST services provided by the platform. Next, during the *normalization* task we obtain a standardized version, according to a lexical resource, of the terms contained in list names. Finally, we *transform* the Twitter list data into a one-mode graph where nodes are the members of the lists, and there exists a weighted edge between two users if they were classified under a list containing a shared term. The weight of each edge corresponds to the number of shared lists.

2.2 Collecting the domain Vocabulary

We traverse the graph starting from some initial users which are prominent in the domain. We propose to use external resources and Twitter information to identify prominent users. External resources can be domain experts or for instance bibliographic resources. Prominent users in real life may not be important in Twitter, so we measure their influence index (e.g., using the klout.com service). The intuition is that around prominent users is more feasible that a vocabulary has emerged since more people are interested in what they are saying. Then we traverse the graph starting from each prominent user and compare the users, that we are reaching while traversing, using the terms under which they have been listed with the terms related to the starting user. The terms of the most similar users are added to the domain vocabulary.

2.3 Eliciting the Vocabulary Semantics

Finally, we use knowledge bases to elicit the semantics of the terms found in the previous activity. We propose to use linked data sets so that we benefit from the interlinked conceptualizations. We associate terms with semantic entities which in turn are used to obtain classes. Next we search for relations between the classes using SPARQL queries. We include relations set up through intermediate semantic entities which can be also included in the final ontology. In addition the ontology can be populated with existing instances of the classes reused in the process. Challenges in this activity include ambiguity of terms and knowledge base heterogeneity. The output is an ontology consisting of classes, relations and instances.

References

1. García-Silva A., Kang J.H., Lerman K and Corcho O. Characterising Emergent Semantics in Twitter Lists. In 9th extended Semantic Web Conference, Crete, 2012 (To appear).