Using Text Mining to Enrich the Vocabulary of Domain Ontologies

Mirco Speretta and Susan Gauch
CSCE Department, University of Arkansas, USA
{msperett, sgauch}@uark.edu

Abstract

Users organize personal information in various ways. We believe that this process could be expedited and improved by using domain ontologies. The main problem with this idea is the lack of automatic tools that help non-expert users to build and maintain their own ontologies. In this study we report progress in the process of adapting ontologies to better represent a given set of documents centered on a topic of interest. More specifically we investigate automatic approaches to enhance the representation of the concepts within the domain ontology. We show that our approach can enrich the vocabulary of each concept with words mined from the set of small documents provided. The method we propose is based on efficient text mining approaches combined with semantic information from WordNet.

1. Introduction

The power of ontologies has not been completely exploited, in part because of the challenge of building and maintaining them. Furthermore, the lack of automatic tools makes these tasks particularly arduous to non-expert users. In a previous study we investigated [1] the scenario in which the user provides a sample documents related to a topic for which an ontological representation is desired. Using these documents, our approach attempts to select the best-matching domain ontology from a large public library. Although it should provide a good start in creating an ontology for a particular user’s interests, the pre-existing ontology is likely to require modifications to better fit the sample documents.

This paper reports on progress in the process of adapting ontologies to better represent a given set of documents. To determine the required adaptations, we focus our attention on the vocabulary coverage of the selected ontology. More specifically, we consider the terminology use in the concept representations and we investigate automatic approaches to enhance the concept representations by mining vocabulary from the sample documents.

In this study, we propose to apply known word relatedness measures to extract single words that can be used to enrich ontologies. To reduce the burden on the user, we want to verify the effectiveness of our technique using a small set of documents that could contain as few as twenty representative Web pages. The method we describe here is based on a combination of text mining techniques with semantic word similarity measures based on the WordNet’ taxonomy.

2. Background

We address the problem of enriching the vocabulary of domain ontologies. In the literature, this problem is better known as taxonomy expansion and it has been studied by many researchers with different perspectives. Faatz and Steinmetz [2] conducted a study based on data collected through a search engine. Mahn and Biemann [3] studied word co-occurrence to extend ontologies. They calculated the significance of a pair of words over sentences or documents.

There are many techniques to measure the semantic relatedness between two words. Budanitsky and Hirst [4] evaluate five measures that are all based on WordNet. They found that the method proposed by Jiang and Conrath [5] was the most effective. Jiang and Conrath’s approach is based on both a taxonomic structure and statistical distribution of words in a corpus. This was an improvement over Resnik’s [6] idea for which the semantic similarity between two concepts depends on the distance with the closest common ancestor concept.

The technique that we propose is very similar to the one investigated by Warin et al. [7]. In that study, the authors used WordNet similarity measures to disambiguate the concepts within an ontology. The approach was quite effective, and we build on their work to enrich the vocabulary coverage of the ontology.

3. Approach

For our experiments, we considered an existing ontology and a set of 100 related Web pages manually selected. The ontology is in the domain of music and it includes 33 concepts, 9 of which are described using 2 words (e.g., “double bass” or “string quintet”) while the remaining 24 are described using a single word (e.g., “cellist” or “composer”). The total number of words in the ontology is 42 while the number of unique words is 32. We applied our approach over what we call the concept-words, which are the words used to label each concept.

1 http://wordnet.princeton.edu/
(e.g., “double bass” is the concept whereas “double” and “bass” are the concept-words). We evaluated four approaches to suggest one or more words to be added to each concept-word.

All experiments were conducted using 5-fold verification in which, in rotation, 80 documents were used for training and 20 for testing. The sample documents were preprocessed using standard techniques to remove html code, stop words, and punctuation. All words from the training set were weighted using tf * idf. The tf value was simply the frequency within the page. The idf value was calculated over a bigger set of 260 documents that includes 9 different topics. This approach allowed us to give a higher rank value to the words that are more representative of our specific domain. We used the top 10% words of the list as potential candidates for the enrichment of the vocabulary.

3.1. Common Hypernyms Approach

This approach is based on the structure of WordNet. From the set of candidates, we select the words that have common hypernyms in WordNet with the concept-words. We consider all the syntactical functions and all the senses of a candidate word (both noun and verb), thus a single concept-word can have multiple hypernyms.

The algorithm works as follows: first, we used WordNet to collect all the hypernyms for each concept-word. Then, we ranked the hypernyms by the number of times each occurred. Finally, to remove noise, we removed the hypernyms that occurred just once. We call the set of remaining hypernyms the common hypernyms.

Since we are interested in the candidate words that are mined from the sample documents and are related to the ontology’s domain, we keep only those with at least one hypernym included in the set of the common hypernyms. Certain candidate words might be associated with more than one concept-word. To avoid introducing ambiguity in the ontology, we pair each candidate word with the associated concept-word with the highest tf * iof weight [1], where iof is a variant of idf that is calculated across our collection of 183 ontologies and increases the weights of tokens that are unique to a particular ontology.

In the case of ambiguity (e.g., a candidate word associated with more than one concept-word with the same tf * iof weight), we pair it with a concept-word chosen randomly. This has the effect of adding the candidates to the concepts with which they are most related.

3.2. Relatedness Measures Approach

In this approach, we exploit semantics by applying existing algorithms that use WordNet to measure the similarity between two given words. First, each concept-word is paired with each candidate word. Then, we calculate the relatedness between the two words by applying a similarity algorithm. Afterwards we sort the list of pairings in decreasing order of similarity values. The list of all pairings is then processed so that each candidate word only appears paired with the concept-word with which it has the highest similarity. We then evaluate the effect of adding a variable number of the highest-ranked candidate words to each concept. We test our approach on the following three similarity algorithms: path between the synsets of two words, Resnik [6], Jiang and Conrath [5]. We used the perl implementation package WordNet::Similarity [8] developed by Ted Pedersen et al. In this study, we are focusing on words representing labels of concepts. In our ontologies, these words are almost exclusively nouns, so, for the relatedness measure approaches, we considered nouns as the part-of-speech for all similarity calculations.

4. Evaluation

The goal of this research study is to enrich the vocabulary of the ontology by associating each concept with semantically related words. The evaluation of this association is non-trivial because it measures the potential correlations between meanings of words. Each approach is evaluated by measuring the quality of the match between the enriched ontology and the test document set represented by 20 Web pages.

We employ three metrics, the first two of which are objective and the third of which is subjective. The document vocabulary coverage measures the percentage of words in the document set that appear in the ontology’s vocabulary whereas precision measures the percentage of words added to the ontology that appear in the document set. Finally, the ontology semantic relatedness measures the correlation between concept-words and the corresponding added words, based on human judgment.

4.1. Document vocabulary coverage

For this evaluation, we measure how well the enriched vocabulary of the ontology covers the given set of documents. This problem can be considered analogous to evaluating a semantic tagger system. Thus, we used the statistical measures applied by Demetriou and Atwell [9]: the vocabulary type coverage and the real text token coverage. We modified these measures into documents vocabulary coverage (DVC) and the real token coverage (RTC). The DVC is an indicator of the proportion of unique words in the document set that appear in the enriched vocabulary whereas RTC measures the proportion of the total words (tokens) in the document set that appear in the ontology’s vocabulary.
4.2. Precision

Since one could maximize the document vocabulary coverage measures by simply adding every word in the training set to the ontology, it is also important to measure how many of the added words actually appear in the test collection. To measure the precision of the new words, we elaborated the approach introduced by Faatz and Steinmetz [10]. For our work, we slightly modified their formula for Precision to:

\[ \text{Precision} = \frac{\left| \{x \in UCW | x \in MW\} \right|}{|UCW|} \times 100 \]

where

- \( UCW \) is the set of unique candidate words
- \( MW \) is the set of matching words

Thus, Precision measures the percentage of the words used to enrich the ontology that are used in the testing set of documents.

4.3. Ontology semantic relatedness

The words added to the ontology might appear in the training and test sets, (e.g., “javascript”), but still be unrelated to the domain of the ontology or they may be added to the wrong concept within the ontology, (e.g., adding Beethoven to ‘instrument’ rather than ‘composer’), The goal of this evaluation is to verify the semantic relatedness between the candidate words and their corresponding concepts. For each concept-word, we presented a native English speaker with the 40 highest-ranked candidate words. She was asked to judge the association on a 3 point scale in which 0 meant “not related”, 1 meant “somewhat related”, and 2 meant “very related”. In total, the judge evaluated a list of 1,368 concept-candidate word pairs.

5. Results

In this section, we present the results from the 4 techniques described in Section 3. For each technique, we assumed that each concept should be enriched with a few highly-related words. Thus, we evaluated adding the top-ranked candidates from a minimum of 1 to a maximum of 40 per concept-word.

5.1. Document vocabulary coverage

In Figure 1, we show the document vocabulary coverage (DTC) as a function of the number of words added per concept. It is not surprising to observe that the coverage of the vocabulary increases proportionally with number of candidate words added. Among the techniques evaluated, the res algorithm performs the best.

Figure 1 Vocabulary Coverage

In Figure 2, we show the real tokens coverage (RTC metric) as a function of the number of candidate words added. We notice similar trends to the document vocabulary coverage in Figure 1. The main difference is that the common hypernyms algorithm performs much worse than the approaches based on word similarity. As with DVC, the res algorithm performs slightly better.

Figure 2 Real Tokens Coverage

5.2. Precision

The Precision metric shows that the path algorithm gives the best results since it shows high Precision for the highest-ranked candidate and then it steadily decrease as less-highly-ranked candidates are added. The res algorithm demonstrates similar behavior after the top 4 candidate words are added. The jcn algorithm does not show a point after which the coverage decreases. The worst performance is given by the common hypernyms algorithm since the Precision tends to increase along with usage of more candidate words, which is counter-intuitive.
5.3. Ontology semantic relatedness

Figure 4 shows the results of the evaluation performed by the judge. She evaluated all candidate words for each algorithm. The x-axis represents the top ranked words considered whereas the y-axis is the average relatedness value provided by the user. We can observe that *jcn*, *path*, and *res* algorithms perform similarly. For each of these approaches, the relatedness is inversely proportional to the rank of the candidates considered.

![Figure 3 Precision Indicator](image)

**Figure 3 Precision Indicator**

From these results, we observe a sudden decrease of relatedness right after the highest-ranked candidate for both the *path* and the *jcn* algorithms. The *path* similarity algorithm is best at identifying the top candidate, but overall the *res* algorithm performs slightly better.

![Figure 4 Semantic Relatedness](image)

**Figure 4 Semantic Relatedness**

6. Conclusions

In this study, we report progress in the process of adapting ontologies to better represent a given set of documents. We accomplish this by investigating automatic approaches to enhance the representation of the concepts within a given domain ontology. We show that our approach can enhance each concept with new candidate words mined from a given set of small documents. We think our findings represent a next step into the feasibility of implementing automatic tools to create and manage ontologies. The method we propose is based on well-known word similarity techniques that can be calculated on the fly. In this way users, both expert and non-expert, could have access to use domain ontologies effectively and automatically for theirs daily activities.

7. Acknowledgement

This work was partially supported by the National Science Foundation under grant NSF DBI-0445752.

8. References


