Grammatical Evolution for Identifying Wikipedia Taxonomies

Lourdes Araujo
Universidad Nacional de Educacion a Distancia (UNED)
Madrid, Spain
lurdes@lsi.uned.es

Juan Martinez-Romo
Universidad Nacional de Educacion a Distancia (UNED)
Madrid, Spain
juaner@lsi.uned.es

Andrés Duque
Universidad Nacional de Educacion a Distancia (UNED)
Madrid, Spain
aduque@lsi.uned.es

ABSTRACT

This work applies Grammatical Evolution to identify taxonomic hierarchies of concepts from Wikipedia. Each article in Wikipedia covers a concept and is cross-linked by hyperlinks that connect related concepts. Hierarchical taxonomies and their generalization to ontologies are a highly useful resource for many applications by enabling semantic search and reasoning. We have developed a system which arranges a set of Wikipedia concepts into a taxonomy.

Categories and Subject Descriptors

I [**Machine Learning Approaches**]: Bio-inspired approaches

Keywords

Grammatical Evolution, Genetic Algorithm, Wikipedia taxonomies

1. THE PROPOSAL

A key step towards the full Semantic Web functionality is the efficient organization of human knowledge in ontologies. These usually large and hand-made structures have to be adapted to new knowledge in an efficient and reliable way.

In this work, we propose a method for automatically organizing parts of a wide spread and constantly updated source of knowledge, which is Wikipedia. Information in Wikipedia is organized in articles, each of them devoted to a particular topic. An interesting question that arises when considering linked Wikipedia pages is the kind of relationship between the linked concepts. In particular we are interested in identifying the "is a" relationship between Wikipedia concepts in order to organize them into a taxonomy or hierarchy.

An example of ontology related to Wikipedia is the DBpedia Ontology [1]. The DBpedia Ontology is a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia. We have used the DBpedia ontology concerning *Species* for evaluation.

In this work we apply grammatical evolution (GE) to discover a function of a set of features extracted from the content of the Wikipedia pages for constructing taxonomies with the associated Wikipedia concepts. The candidate function being evaluated should approximate the hierarchical relationships between the concepts of each part of the *Species* ontology. The fitness is computed as the average for the set of training taxonomies, of the precision obtained when comparing the taxonomy with the highest score according to the function, with the reference one. In order to obtain the highest score taxonomy that a candidate function can provide, we need to perform an optimization process which is, in turn, implemented by a genetic algorithm. Figure 1 shows a scheme of the system. Wikipedia provides the linked pages of articles related to a set of concepts. From the terms contained in each of these documents we compute a weighted term vector associated to the corresponding concept. Different relationships can be expected to be fulfilled between the vectors associated to related concepts. Then, a function that appropriately combines these features can detect the hierarchical relation between two concepts. The grammatical evolution algorithm evolves functions combining the considered features. The fitness of a candidate function is computed by comparing the approximate best taxonomy that the function can obtain for a number of training cases, with the taxonomy of reference of each case. The best taxonomy for a function and training case is obtained applying a genetic algorithm, which uses the term vectors representing the documents to compute the value of the features appearing in the function.

2. THE WIKIPEDIA TAXONOMIES PROBLEM

We adopt the Vector Space Model [3] for representing Wikipedia pages. In this model text documents are represented as vectors of terms, where the value of each term ($w_{i,j}$) indicates the relevance of the term as representative of the document d_j. We weights each term with the TF-IDF (term frequency - inverse document frequency) measure, where TF, $tf(t,d)$, stands for the frequency of a term t in a document d, and IDF, $idf(t,d)$, for the inverse document frequency of a term t along all the documents d in the considered collection D.

The set of relationships considered, which tend to be met between two linked pages (i.e. their corresponding vectors) with a hierarchical relationship are:

GECCO '15 July 11-15, 2015, Madrid, Spain

ACM ISBN 978-1-4503-3488-4/15/07.

DOI: http://dx.doi.org/10.1145/2739482.2764629

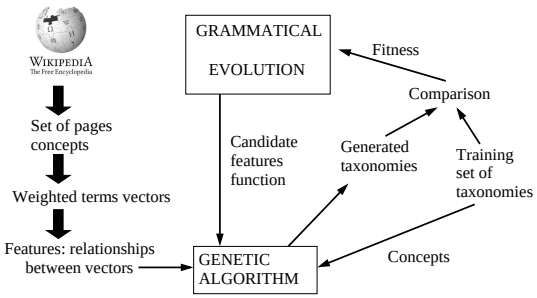

Figure 1: System scheme. The input data to the algorithm are the Wikipedia pages associated to different concepts, and the training set of taxonomies. The Grammatical Evolution algorithm uses a Genetic Algorithm to compute the fitness of the candidate functions.

- **COS (cosine):** The most popular similarity measure is the cosine coefficient, which measures the angle between the two document vectors. It is commonly used to detect if two documents are really related to each other.

- **DIFSIM (Differences in similarity):** This measure gives an approximation to the similarity between the intersection S of two vectors A and B, and any of the vector, A or B.

- **Distinct terms**: Specifically we have considered the following measures related to the distinct terms, i.e. terms non-shared by the two concepts:

  - **Average Weight of Distinct terms (AWD)**: the average weight of terms appearing in one concept but not in the other one.

  - **Standard Deviation of Distinct terms (SDD)**: the standard deviation of the weight of distinct terms in each concept.

  These features tend to adopt higher values for the Parent-Child relationship than for the Child-Parent one.

We apply Grammatical Evolution (GE) [2] to combine these features in a function able to detect the tendency of Wikipedia linked concepts to present a hierarchical relationship.

The BNF grammar (Figure 2) has been designed so as to include the features that have been identified as indicators of possible hierarchical relationships: cosine similarity (COS), the difference in the similarity of each concept and the intersection of both (DIFSIM), and the relevance of the distinct terms (AWD) and their deviation (SDD).

The fitness of the GE algorithm is computed as the average precision achieved by comparing, for a number of training cases, the taxonomy provided by the candidate function and the reference taxonomy. In order to obtain the taxonomy which optimizes the value of the candidate function for a particular set of concepts we have resorted to a genetic algorithm.

```
<expr> ::= <op> <var> <var>
                | if <cond> <expr>
                | <var>
<op> ::= + | - | / | *
<cond> ::= <var> = <var> | <var> < <var>
    | <var> > <var> | <var> >= <var>
    | <var> <= <var> | <var> = <cst>
    | <var> < <cst> | <var> > <cst>
    | <var> >= <cst> | <var> <= <cst>
<var> ::= COS | DIFSIM | AWD1
                | AWD2 | SDD1 | SDD2
<cst> ::= 0.05 | 0.1 | ... | 0.9 | 1 | ... | 5
```

Figure 2: BNF Grammar for the algorithm.

Table 1: Best functions found.

ID	Function
F1	expr(if(cond(var(DIFSIM) > cst(0.3)), expr(op(/),var(AWD1),var(AWD2))))
F2	expr(if(cond(var(AWD2) > cst(0.3)), expr(op(/),var(AWD1),var(AWD2))))
F3	expr(if(cond(var(AWD2),>,var(DIFSIM)), expr(op(-),var(AWD1),var(SDD2))))

2.1 Results and Discussion

Table 1 shows some of the best functions found by the algorithm along different runs. The GE algorithm has been able to provide valuable functions that combine the considered features extracted from the Wikipedia pages. These functions have been able to correctly identify some part of a possible taxonomy among Wikipedia concepts, such as *Animal* and *Plant*. Even in the cases in which the obtained taxonomy does not match the DBPedia taxonomy used as reference, we can see that the method has been able to detect real relationships such as the ones between *Insect* and *Arachnid*, and *Crustacean* and *Fish*. Best results are obtained between groups of concepts which are directly connected in Wikipedia. Results get worse for the most general concepts, such as *Species*, the top of the considered taxonomy. However, other features can be extracted from the Wikipedia pages and the proposed algorithm can be used to find the best function to combine them.

3. ACKNOWLEDGMENTS

This work has been partially supported by the Universidad Nacional de Educación a Distancia (UNED) within the project 2013-025-UNED-PROY, and the Spanish Ministry of Science and Innovation within the project EXTRECM (TIN2013-46616-C2-2-R).

4. REFERENCES

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[2] M. O'Neill and C. Ryan. Grammatical evolution. *IEEE Transactions on Evolutionary Computation*, 5(4):349–358, August 2001.

[3] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.