Aligning Multi-Cultural Knowledge Taxonomies by Combinatorial Optimization

Natalia Prytkova
Max-Planck-Institut für Informatik
Saarbrücken, Germany
natalia@mpi-inf.mpg.de

Marc Spaniol
Université de Caen Basse-Normandie
Caen Cedex, France
marc.spaniol@unicaen.fr

Gerhard Weikum
Max-Planck-Institut für Informatik
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Large collections of digital knowledge have become valuable assets for search and recommendation applications. The taxonomic type systems of such knowledge bases are often highly heterogeneous, as they reflect different cultures, languages, and intentions of usage. We present a novel method to the problem of multi-cultural knowledge alignment, which maps each node of a source taxonomy onto a ranked list of most suitable nodes in the target taxonomy. We model this task as combinatorial optimization problems, using integer linear programming and quadratic programming. The quality of the computed alignments is evaluated, using large heterogeneous taxonomies about book categories.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: Models and Principles

Keywords

Multicultural Knowledge Taxonomies; Alignment Methods; Integer Linear Programming; Quadratic Programming

1. INTRODUCTION

Large knowledge bases (KB's) have gained much attention, as they are powering Internet and enterprise search. Major success stories include knowledge graphs at Google, Microsoft, Walmart, Bloomberg and others, as well as academic projects such as DBpedia, NELL, YAGO, etc. A crucial backbone for proper interpretation of this rich knowledge are *taxonomies*: tree- or DAG-shaped hierarchies of semantic types or thematic topics to which entities are assigned. In addition to these KB's in a strict sense, there are numerous knowledge collections, e.g., product catalogs such as amazon.com, digital libraries such as the US Library of Congress or the German National Library, Wikipedia editions, specialized online communities on health issues, music, etc.

Together all this constitutes a "knowledge habitat" of complementary information from different domains, cultures, and languages. Therefore, there is an inherent need to "translate" the contents from one taxonomic structure into another. However, making a transition from a category in one taxonomy to the most relevant counterpart(s) in another is not a trivial task.

For example, category `Kinder- & Jugendliteratur` in amazon.de (en. Children & Youth Literature) has two relevant counterparts in amazon.com – `Children's Books` and `Teen & Young Adult`. Neither a translation tool is able to resolve the semantic equivalence of the categories nor can the topological comparison of taxonomic trees accomplish this.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742721.

2. BASIC METHOD

For taxonomies K_1 and K_2, we compute a baseline alignment in two steps, harnessing the category system of the English Wikipedia as an intermediate taxonomy:

1. Compute semantic labels for all nodes i and j of K_1 and K_2, respectively, via mappings to the Wikipedia taxonomy, using either the overlap of instances of i and j or the string similarity of i and j to the names of Wikipedia categories.
2. Generate candidate mappings between K_1 and K_2 by considering all pairs i, j that share semantic labels.

3. ADVANCED METHODS

The basic alignment maps each source category $i \in K_1$ to a set of candidate targets $j_1, j_2 ... \in K_2$ in isolation and it does not consider the correlation between the candidates.

We model a correlation-aware mapping as a combinatorial optimization problem. For each pair of categories $i \in K_1$ and $j \in K_2$, which share at least one semantic label, we create a binary variable $A_{i,j}$. $A_{i,j}$ is set to 1 if categories i and j are aligned in the current solution. Otherwise, it is 0.

We consider two optimization models: i) a model with hard constraints, using integer linear programming (ILP), and ii) a model with soft constraints, using quadratic programming (QP). Both of these can be solved by standard tools like Gurobi; our contribution is the novel way of modeling.

4. HARD CONSTRAINTS IN ILP

The goal is to find an alignment with the maximal weight, which is expressed as the following objective function:

$$\max \sum_{i \in K_1, j \in K_2} w(i,j) \cdot A_{i,j} \qquad (1)$$

where $w(i, j)$ is the weight of alignment between i and j based on shared semantic labels.

This objective is subject to the following constraints that ensure the coherence of the target candidates by eliminating non-correlated candidates.

$$\begin{aligned} A_{i,j} + A_{i,k} &\leq 1 \text{ if } corr(j,k) \leq 0 \\ A_{i,j} + A_{u,j} &\leq 1 \text{ if } corr(i,u) \leq 0 \end{aligned} \quad (2)$$

where $corr(x, y)$ is the Pearson's correlation coefficient between entity vectors of the categories. Entity vectors capture the occurrence frequency of entities in categories.

In theory, the number of constraints is cubic. However, in practice their number is much smaller since (a) the number of A variables is limited to those pairs which are connected via an intermediate category and (b) constraints are added only for the non-correlating target pairs. For example, the model for the amazon.de → amazon.com case has 10^7 constraints, whereas in theory it should have had about 7×10^{11}.

5. SOFT CONSTRAINTS IN QP

Forcing all candidate targets to be positively correlated may be too aggressive. Instead, we can relax the anti-correlation constraint and define a "soft" variant by a reward term in the objective function of the combinatorial optimization.

For source category i and candidate targets $j_1, j_2...$ the reward is the pairwise correlation between all target categories. We denote this by $corr_{K_2}$:

$$corr_{K_2} = \sum_{i \in K_1} \sum_{j \in K_2} \sum_{k \in K_2} corr(j,k) \cdot A_{i,j} \cdot A_{i,k} \quad (3)$$

Analogously, we can define the reward for pairwise correlation of the source categories that would be aligned with the same target. We denote this as $corr_{K_1}$.

We extend the objective function, beyond merely maximizing the alignment weight, by maximizing the sum of the alignment weight and the two reward terms. The objective function of this model becomes:

$$\max[\sum_{i \in K_1, j \in K_2} w(i,j) \cdot A_{i,j} + corr_{K_1} + corr_{K_2}] \quad (4)$$

Note that the reward terms have a product of decision variables, resulting in a quadratic optimization model.

	amazon.de → amazon.com	shelfari.com → dnb.de
baseline	0.78 ± 0.08	0.85 ± 0.06
ILP	0.48 ± 0.10	1.00 ± 0.00
QP	0.81 ± 0.08	0.93 ± 0.05

Table 1: Experimental results: MRR values

6. EXPERIMENTAL EVALUATION

To evaluate the quality of the alignments computed by our methods, we performed experiments with different taxonomies and human judges for assessment. We use four taxonomies:

- *amazon.com* and *amazon.de* - English and German product catalogs, which are independent of each other and have different category systems (for English: 5,846 categories, 28,754 authors, 1,724,943 books; for German: 8,293 categories, 28,360 authors, 933,779 books);
- *shelfari.com* - community-created category system of books (12,803 categories, 559,877 authors, 1,159,897 books)
- *dnb.de* - categorization of the books in the German National Library (910 categories, 421,896 authors, 751,346 books).

In all of these, the entities of interest are books and authors. As intermediate taxonomies we use the English and German Wikipedia editions with their categories as semantic labels.

Since there is no ground-truth to compare with, we manually evaluated the quality of the generated alignments for a randomly generated sample of 100 source categories, for each of the taxonomy pairs. Two judges annotated each pair of categories from the alignment output of each method as *wrong* or *correct*. For each source category we allowed at most one target category to be labelled as *correct*.

Experimental results are given in Table 1. We report the mean reciprocal rank (MRR) of the correct target in the candidate list.

The baseline performs surprisingly well. ILP is superior to the baseline on one of the two presented cases. QP is the overall winner with very good results on MRR in both test cases. QP treats the correlations between categories in a more elegant way than ILP. If an erroneous target is assigned to a source, then by hard constraints, all possibly matching target are removed from the candidate list. This explains, why ILP does not perform well in one of the use cases. QP relaxes the constraints and is, thus, more flexible.

As an example, QP computed a pair of equivalent categories `Medizin/Innere Medizin` in amazon.de and `Medicine/Internal Medicine` in amazon.com, whereas the baseline and ILP solutions returned the wrong match `Veterinary Medicine/Cardiology`.

7. RELATED WORK

Data integration ([2]) computes mappings between mediation schemas and local database schemas. This setting is quite different from aligning culturally diverse taxonomies. **Ontology alignment** ([3, 7, 8]), deals with full-fledged ontologies, whereas we concentrate on taxonomic structures. In our setting, the number of taxonomic categories is orders of magnitude larger than the number of schema elements that ontology alignment methods can handle. **Multilingual data and knowledge alignment** generates missing links across Wikipedia editions [4], or interlinks Wikipedia infoboxes from different languages [6]. Our work addresses a much wider variety of taxonomies beyond Wikipedia. **Catalog integration** aims at finding similar categories in Internet directories ([1, 5]). In contrast to these works, we do not assume any direct mapping of instances in different taxonomies, reflecting the cultural diversity in our setting.

8. REFERENCES

[1] R. Agrawal, R. Srikant. On Integrating Catalogs. *WWW* 2001.

[2] A. Doan, A. Halevy, Z. G. Ives. *Principles of Data Integration*. 2012.

[3] J. Euzenat, P. Shvaiko. *Ontology Matching*. 2013.

[4] J. Göbölös-Szabo et al. Cross-Lingual Data Quality for Knowledge Base Acceleration Across Wikipedia Editions. *QDB Workshop* 2012.

[5] R. Ichise et al. Integrating Multiple Internet Directories by Instance-Based Learning. *IJCAI* 2003.

[6] T. Nguyen et al. Multilingual Schema Matching for Wikipedia Infoboxes. *PVLDB* 2011.

[7] O. Udrea, L. Getoor, R. J. Miller. Leveraging Data and Structure in Ontology Integration. *SIGMOD* 2007.

[8] M. L. Wick et al. A Discriminative Approach to Ontology Mapping. *NTII* 2008.