Identifying constitutive articles of cumulative dissertation theses by bilingual text similarity. Evaluation of similarity methods on a new short text task

Paul Donner

Appendix

A1 Heuristic candidate selection

In order not to compare a thesis record with all Scopus records we developed a candidate selection method which discards very unlikely Scopus records while retaining any reasonably plausible records. We filter Scopus records by author name match, country, and publication dates.

An author name table was created for all authors present in Scopus with German address data who published any of the relevant document types article, review, conference paper, book chapter or book. Normalized variants of names were added to increase matching by removing non-letter characters, including all parts of double-barreled family names, including alternatives for German letters ß, Ä, Ö, Ü. Thesis author names were then checked against Scopus record author names with Jaro-Winkler modified similarity function versions adapted for names such as to take into account initials, capitalizing all letters, and simplifying all letters to ASCII letters. Candidate matches between theses and Scopus records were retained for experimentally chosen similarity thresholds and filtered for a publication year range from four years before thesis publication to nine years after.

We noticed a considerable number of names in both the thesis data and in Scopus where given name and family name were obviously switched. In order to find these cases, a routine was run which splits all names into parts on both datasets and finds additional matches irrespective of name part order. In other words, the name matching was done without requiring that family name match family name or given name match given name. Thesis and Scopus name pairs for which two or more name parts could be matched were added to the candidate matches if the publication year condition also applied.

No candidate articles were identified for 5 doctoral thesis records, all of them without cumulative articles according to inspection. The maximum found candidates for one thesis is 47,654. On average there are 1464 candidates per thesis record.

This filtering step failed to identify a total of 48 thesis-associated articles. The following causes were identified:

*German Centre for Higher Education Research and Science Studies (DZHW), Dept. 2 ‘Research System and Science Dynamics’, Schützenstr. 6a, 10117 Berlin, Germany. donner@dzhw.eu. ORCID 0000-0001-5737-8483
1. Seven articles due to document types not deemed relevant, such as editorials and notes.
2. Changed family names led to 24 articles of women not being found.
3. One item was not identified because the thesis author was not indexed in Scopus as author of a book chapter which he co-authored.
4. Eight items were not found because the name formats of complex multi-part author names of the DNB and Scopus data were too divergent.
5. Five items of two thesis authors were not identified because they had no German affiliations in their Scopus-indexed articles.
6. Three articles of a single thesis author were not found because the differences of publication years between thesis and articles were too great (16, 17, and 19 years).

Items 1 and 3 are not due to the selection method. Leaving these aside, the method failed for 40 out of 1495 pairs, which is an accuracy of over 97%. Items 4, 5, and 6 could potentially be overcome by changing the inclusion criteria but this would lead to a disproportionate amount of further negative candidates.

A2 Detailed results for parametrized methods

In this section we present further detailed results for the LSA, RI, and RRI methods. These methods have one parameter, the dimensionality of the created vector space, which was tested in the range 100 to 1000 by increments of 100. Figure S1 shows the recall curves for ranks up to 20 for theses with associated articles. For LSA we observe a regular increase in performance with increasing dimensionality. For RI, low-dimension vector spaces also perform less well but the gains are not regular and the models with dimensions from 700 to 1000 perform equally well. The different RRI models, in contrast, all perform indistinguishably by this evaluation method, no effect of the vector space size can be found. Figure S2 shows a subset of these data from a different perspective. The x-axis gives the dimensionality parameter of the models and the y-axis the recall at rank 20. Again we see regular gain in performance of LSA with increasing vector space size. Gains for RI are only visible at lower dimensions. RRI seems to perform at the same level independent of the dimensionality parameter.

Figure S3 shows the average standardized similarity scores for two groups of values per method. The first group, symbol ‘-‘ in the figure, are the highest similarity values of articles for those theses records without any associated articles. The other group (‘+’ symbol) are the mean scores of the true positives, that is, the known elemental articles of cumulative dissertation theses, for those records with such articles. The differences for the parametrized models across different vector space sizes are minor. The ability of the LSA method to globally differentiate true positives from other similar records seems to decrease with dimensionality, that of RI slightly increases, while that of RRI remains unchanged. For RI 100, the difference between average true positive and other standardized similarities is about -0.1 SD, while for RI 1000 the difference is about 0.3 SD.
Figure S1. Recall across similarity rank positions, LSA and RI/RRI, all dimensionalities
Figure S2. Average recall at rank 20, LSA, RI, RRI, all dimensionalities
Figure S3. Differences in standardized similarity scores