Semantic knowledge discovery

Gustaf Nelhans*, Johan Eklund
*Corresponding author | name.lastname@hb.se

Data as Impact Lab, Swedish School of Library and Information Science, University of Borås

Since many databases lack relevance ranking, a citation-based approach can be a valuable complement since it is possible to use citation-based data to indicate centrality, relevance, or visibility in the research community. However, using bibliometric methods in the humanities is often challenging since a lot of the research literature is not indexed in the traditional citation databases that we generally use for bibliometric mapping.

We introduce a combined bibliometric and semantic approach to extend a network of bibliographic records by incorporating a larger set of records lacking bibliometric features based on the semantic similarities between their titles. In order to expand the set of identified relevant articles, we used the Universal Sentence Encoder (USE) algorithm developed by Google Research to generate semantic vectors for the titles.

We searched several different databases, of which some include citation data, to create a pool C of candidate documents within the selected subject area. A set A of documents was obtained from a citation database to generate the initial network of articles. We then calculated the bibliographic coupling of articles as quantified by their shared references.

We manually selected a small set $S_1 \subset A$ of documents representing different topical clusters as a seed for the expansion based on semantic similarities. For each document $d \in S_1$, we ranked the documents in C in ascending order according to their cosine distance to the title vector assigned to d, then selecting the k documents closest to d. This procedure gave us a set $S_2 \subset C$ of documents to read.

The results were evaluated using qualitative analysis to determine they were thematically relevant to the present information needs.