zhang_2018_does_deep_learning_help_topic_extraction_a_kernel_k_means_clustering_method_with_word_embedding

Year

2018

Author(s)

Yi Zhang and Jie Lu and Feng Liu and Qian Liu and Alan Porter and Hongshu Chen and Guangquan Zhang

Title

Does deep learning help topic extraction? A kernel k-means clustering method with word embedding

Venue

Journal of Informetrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Novel approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

A two-round expert evaluation was designed and seven leading bibliometric experts were engaged in the evaluation.

Round 1 followed the way of traditional questionnaires to invite experts to mark the grouped topics in Table 5, and three criteria were raised.

Each expert was asked to score the three criteria.

Generally, 1 meant excellent agreement, 0 meant strong disagreement, while an intermediate judgment (e.g., 0.7) was fine as well.

- Coherence: How well do the terms of the Topic go together?
- Distinctiveness: Is the Topic separate from the others?
- Significance: Is the Topic important within bibliometrics? But note that this is an extra task for topic extraction, since we focus on clustering rather than identifying emerging topics.

Two leading bibliometric experts participated in this evaluation, and the average scores of their evaluation results and their correlation coefficient are presented in Table 6.

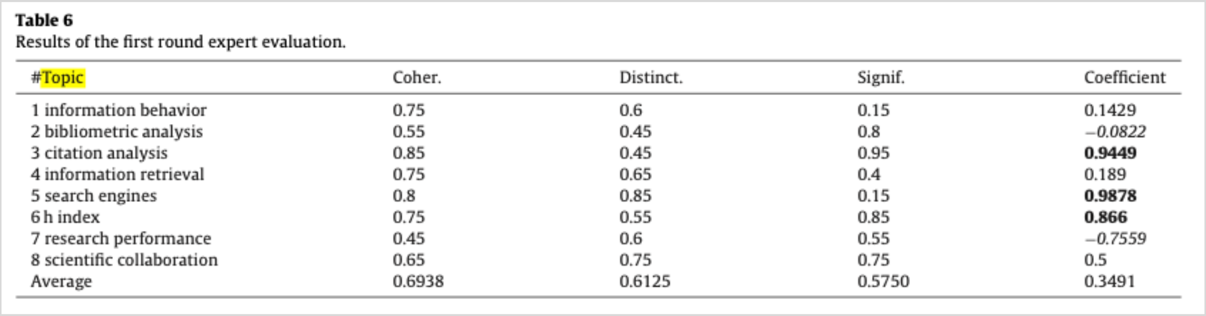

Table 6
Results of the first round expert evaluation.

#Topic	Coher.	Distinct.	Signif.	Coefficient
1 information behavior	0.75	0.6	0.15	0.1429
2 bibliometric analysis	0.55	0.45	0.8	*−0.0822*
3 citation analysis	0.85	0.45	0.95	**0.9449**
4 information retrieval	0.75	0.65	0.4	0.189
5 search engines	0.8	0.85	0.15	**0.9878**
6 h index	0.75	0.55	0.85	**0.866**
7 research performance	0.45	0.6	0.55	*−0.7559*
8 scientific collaboration	0.65	0.75	0.75	0.5
Average	0.6938	0.6125	0.5750	0.3491

In Round 2, five bibliometric experts (different from the ones in Round 1) were involved. Besides Coherence, we raise a criterion of Relevance, i.e., is the Topic relevant with your own research (i.e., bibliometrics)?

We mixed up the 40 core bibliometric terms in Table 5 and asked experts to come up with N clusters (they can decide the N) based on their expertise. Each cluster should represent an area of research in bibliometrics (in some cases, information & library sciences), and each term can only be used once. After that, the experts would consider their own research and interest, and score these clusters, in which 1 meant excellent relevance; 0 meant strong irrelevance, and an intermediate judgment is acceptable as well. This evaluation provides a relatively fair way to generate golden standards for the topic evaluation.

We then set the topics given by the five experts as the golden standards respectively and evaluated our generated eight topics as follows (similar with the JC index):

1. each topic contains five descriptive terms, i.e., 10 distinct term pairs;
2. We then looked for the pairs in the golden standards and confirmed whether the two terms of a pair are within one topic or not. If so, we consider this pair is grouped correctly
3. The percentage of correct pairs in the 10 distinct pairs of each topic is considered as the performance of our proposed method.

The results of the Round 2 evaluation are given in Table 7.

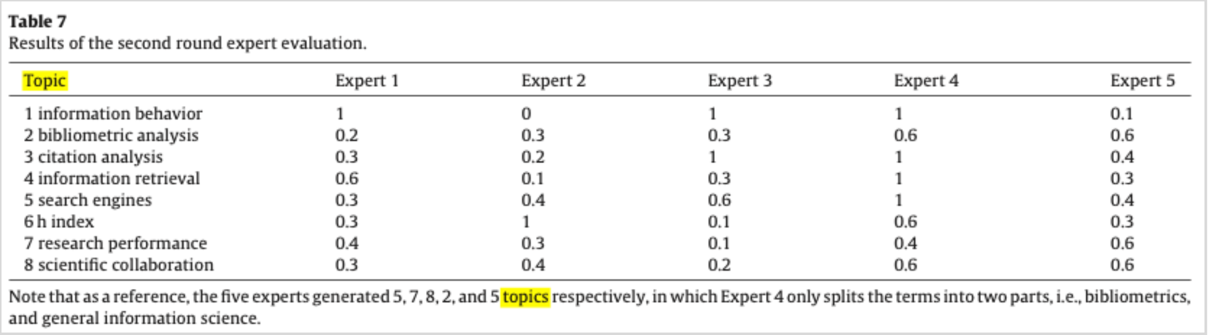

Table 7
Results of the second round expert evaluation.

Topic	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
1 information behavior	1	0	1	1	0.1
2 bibliometric analysis	0.2	0.3	0.3	0.6	0.6
3 citation analysis	0.3	0.2	1	1	0.4
4 information retrieval	0.6	0.1	0.3	1	0.3
5 search engines	0.3	0.4	0.6	1	0.4
6 h index	0.3	1	0.1	0.6	0.3
7 research performance	0.4	0.3	0.1	0.4	0.6
8 scientific collaboration	0.3	0.4	0.2	0.6	0.6

Note that as a reference, the five experts generated 5, 7, 8, 2, and 5 topics respectively, in which Expert 4 only splits the terms into two parts, i.e., bibliometrics, and general information science.

Motivation

Using labelling activities (and related results) as a proxy to qualitatively evaluate the proposed method:

- "it is interesting to qualitatively evaluate the method's performance in a specific domain, with the aid of leading domain experts."
- "Aiming to further evaluate the ability of our proposed method on topic extraction, …"

Topic modeling

K-means clustering method
Baselines: k-means algorithm, a principal component analysis (PCA) algorithm, a topic modeling algorithm, and a fuzzy c-means algorithm

Topic modeling parameters

Nr of topics (k): 8
Kernel function: polynomial

Nr. of topics

8

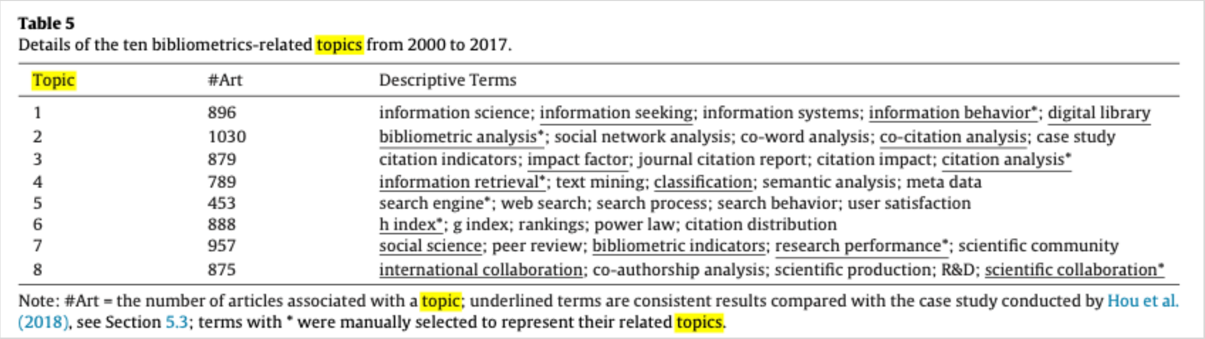

Table 5
Details of the ten bibliometrics-related topics from 2000 to 2017.

Topic	#Art	Descriptive Terms
1	896	information science; information seeking; information systems; information behavior*; digital library
2	1030	bibliometric analysis*; social network analysis; co-word analysis; co-citation analysis; case study
3	879	citation indicators; impact factor; journal citation report; citation impact; citation analysis*
4	789	information retrieval*; text mining; classification; semantic analysis; meta data
5	453	search engine*; web search; search process; search behavior; user satisfaction
6	888	h index*; g index; rankings; power law; citation distribution
7	957	social science; peer review; bibliometric indicators; research performance*; scientific community
8	875	international collaboration; co-authorship analysis; scientific production; R&D; scientific collaboration*

Note: #Art = the number of articles associated with a topic; underlined terms are consistent results compared with the case study conducted by Hou et al. (2018), see Section 5.3; terms with * were manually selected to represent their related topics.

Label

Multi-word label representing an area of research in bibliometrics.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Bibliometrics (topic extraction)

Dataset: Bibliometrics

Problem statement

This paper proposes a novel kernel k-means clustering method incorporated with a word embedding model to create a solution that effectively extracts topics from bibliometric data.
The experimental results of a comparison of this method with four clustering baselines (i.e., k-means, fuzzy c-means, principal component analysis, and topic models) on two bibliometric datasets demonstrate its effectiveness across either a relatively broad range

of disciplines or a given domain.

Corpus

Origin: Web of Science database
Nr. of documents: 6767
Details:

- 3359 SCIM articles, 2784 JASIST articles, and 668 JOI articles, which was further narrowed to 6767 articles that contained both a title and an abstract.

Document

Title and abstract of a scientific article from WoS (SCIM / JASIST / JOI)

Pre-processing

1. Removing terms starting with non-alphabetic characters (e.g., “1.5%”)
2. Removing meaningless terms (e.g., pronouns, prepositions, and conjunctions)
3. Removing common terms in scientific articles (e.g., “method”)
4. Consolidating terms with the same stem (e.g., singular/plural words)
5. Removing terms appearing in only one article
6. Removing single words (e.g., “dataset”).

```
@article{zhang_2018_does_deep_learning_help_topic_extraction_a_kernel_k_means_clustering_method_with_word_embedding,
  abstract = {Topic extraction presents challenges for the bibliometric
community, and its performance still depends on human intervention and its
practical areas. This paper proposes a novel kernel k-means clustering method
incorporated with a word embedding model to create a solution that effectively
extracts topics from bibliometric data. The experimental results of a
comparison of this method with four clustering baselines (i.e., k-means, fuzzy
c-means, principal component analysis, and topic models) on two bibliometric
datasets demonstrate its effectiveness across either a relatively broad range
of disciplines or a given domain. An empirical study on bibliometric topic
extraction from articles published by three top-tier bibliometric journals
between 2000 and 2017, supported by expert knowledge-based evaluations,
provides supplemental evidence of the method's ability on topic extraction.
Additionally, this empirical analysis reveals insights into both overlapping
and diverse research interests among the three journals that would benefit
journal publishers, editorial boards, and research communities.},
  author = {Yi Zhang and Jie Lu and Feng Liu and Qian Liu and Alan Porter and
Hongshu Chen and Guangquan Zhang},
  date-added = {2023-03-15 15:58:03 +0100},
  date-modified = {2023-03-15 15:58:03 +0100},
  doi = {https://doi.org/10.1016/j.joi.2018.09.004},
  issn = {1751-1577},
  journal = {Journal of Informetrics},
  keywords = {Bibliometrics, Topic analysis, Cluster analysis, Text mining},
  number = {4},
  pages = {1099-1117},
  title = {Does deep learning help topic extraction? A kernel k-means
clustering method with word embedding},
  url = {https://www.sciencedirect.com/science/article/pii/S1751157718300257},
  volume = {12},
  year = {2018}}
```

#Thesis/Papers/Initial