kim_2022_developing_a_topic_driven_method_for_interdisciplinarity_analysis

Year

2022

Author(s)

Hyeyoung Kim and Hyelin Park and Min Song

Title

Developing a topic-driven method for interdisciplinarity analysis

Venue

Journal of Informetrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

Nr of examined keywords: 15

Label generation

Prior to labeling, our researchers reviewed the classification of literacy research domains in the literature.
In previous studies, the literacy domain has been divided into “digital literacy, information literacy, data literacy, media literacy, early literacy, health literacy, and physical literacy”.

Thereafter, labels were attached to each topic. The labeling guidelines for each topic were established as follows.

1. First, among the keywords listed in the order of probability values in Table 4, the keyword with the highest probability value was given priority.
2. Second, when the top-ranked keywords overlapped, the labeling was attached to comprehensively reveal the meaning of the topic taking into account the correlation with the topic’s other keywords.

Table 4
15 topics and keywords from DMR topic modeling.

Topic	Labeling	15 Keywords
T1	Social network & education	social networks, science education, social network, sustainability, privacy, ICT, digital literacy, scientific literacy, online learning, climate change, motivation, information technology, social networking, environmental education, trust
T2	Early literacy	academic achievement, parenting, kindergarten, early childhood, early childhood education, motivation, preschool, executive function, school readiness, intervention, elementary school, parents, middle school, early literacy, leadership
T3	Game & e-learning literacy	online learning, collaborative learning, games, game theory, virtual reality, design, video games, multimedia, game design, blended learning, serious games, mobile learning, digital literacy, e-learning, language learning
T4	Public health literacy	diabetes, adherence, depression, type 2, older adults, diabetes mellitus, hypertension, healthcare, medication adherence, HIV, health education, quality of life, nutrition, obesity, public health
T5	Information literacy	theory, e-commerce, cybernetics, information technology, design, computer-mediated communication, ICT, Facebook, IT, e-government, China, information, languages, history, identity
T6	Early language literacy	phonological awareness, dyslexia, kindergarten, vocabulary, music, intervention, speech, comprehension, early literacy, language learning, spelling, emergent literacy, working memory, language impairment, discourse
T7	Digital & collaborative learning	ICT, science education, early childhood education, collaborative learning, pedagogy, teachers, educational technology, digital literacy, interactive learning environments, online learning, media literacy, early childhood, case study, adult learning, identity
T8	Health information literacy	cancer, patient education, healthcare, readability, health information, breast cancer, informed consent, asthma, adherence, depression, nursing, decision making, medical informatics, health education, public health
T9	Democracy literacy	identity, discourse analysis, politics, women, religion, globalization, discourse, history, youth, leadership, language policy, democracy, rhetoric, migration, ethnography
T10	Environment & financial literacy	sustainability, climate change, financial literacy, science education, environmental education, scientific literacy, engineering education, citizen science, sustainable development, climate, youth, biodiversity, financial education, environmental literacy, high school
T11	Language obstacles literacy	dyslexia, working memory, phonological awareness, executive function, speech, developmental dyslexia, brain, autism, auditory processing, attention, phonology, cognition, deaf, intervention, speech perception
T12	Schools & teachers	identity, pedagogy, science education, discourse analysis, social justice, middle school, discourse, teachers, secondary school, diversity, teacher, teacher preparation, multimodal, equity, high school
T13	Library & computer literacy	academic libraries, libraries, financial literacy, library, information technology, librarians, nursing, information, computer literacy, library instruction, computer-mediated communication, information seeking, public libraries, library services, students
T14	Methods & materials	comprehension, methods and materials, digital, media literacies, instructional strategies, teaching/learning strategies, theoretical perspectives, engagement, 4-adolescence, 2-childhood, 3-early adolescence, strategies, motivation, instructional strategies, methods and materials, content literacy
T15	Health literacy for women	cancer, women, breast cancer, cervical cancer, HIV, pregnancy, obesity, depression, public health, health education, cancer screening, health promotion, screening, fertility, colorectal cancer

Motivation

1. Identifying the the literacy domain associated with a topic and
2. Replacing the topic keywords with the associated literacy domain (i.e. label) in order to measure interdisciplinarity in the research domain of literacy

Table 5
Interdisciplinarity indices by topic.

Topic	Label	Variety	Balance	Disparity	Diversity
T1	Social network & education	181	0.335	**0.236**	0.614
T2	Early literacy	178	0.346	0.241	**0.620**
T3	Game & e-learning literacy	**189**	**0.328**	0.263	0.617
T4	Public health literacy	**195**	**0.328**	**0.304**	**0.620**
T5	Information literacy	181	0.351	0.273	**0.609**
T6	Early language literacy	171	0.358	**0.232**	**0.600**
T7	Digital & collaborative learning	178	0.351	0.246	**0.603**
T8	Health information literacy	**190**	**0.326**	0.263	0.613
T9	Democracy literacy	185	0.333	0.246	0.614
T10	Environment & financial literacy	184	0.338	0.254	0.617
T11	Language obstacles literacy	180	0.341	**0.238**	0.617
T12	Schools & teachers	187	0.345	**0.290**	0.610
T13	Library & computer literacy	187	**0.326**	0.242	0.614
T14	Methods & materials	**137**	**0.473**	**0.297**	**0.605**
T15	Health literacy for women	176	0.364	0.268	**0.627**

3. Constructing a network among the 15 nodes and identifying clusters for a more detailed topic-based interdisciplinary analysis

Table 6
Topic-based interdisciplinarity values by six clusters.

Cluster	Topic	Label	Variety	Balance	Disparity	Diversity	DegreeCentrality	BetweennessCentrality
#1	T4	Public health literacy	**195**	**0.328**	**0.304**	**0.620**	10.000	3.400
	T8	Health information literacy	**190**	**0.326**	0.263	0.613	11.000	2.833
	T15	Health literacy for women	176	0.364	0.268	**0.627**	10.000	4.400
#2	T6	Early language literacy	171	0.358	**0.232**	**0.600**	11.000	8.329
	T2	Early literacy	178	0.346	0.241	**0.620**	11.000	11.033
	T11	Language obstacles literacy	180	0.341	**0.238**	0.617	6.000	0.000
#3	T14	Methods & materials	**137**	**0.473**	**0.297**	**0.605**	4.000	9.178
#4	T1	Social network & education	181	0.335	**0.236**	0.614	17.000	13.571
	T3	Game & e-learning literacy	**189**	**0.328**	0.263	0.617	7.000	1.282
	T10	Environment & financial literacy	184	0.338	0.254	0.617	10.000	2.889
#5	T5	Information literacy	181	0.351	0.273	**0.609**	10.000	5.250
	T13	Library & computer literacy	187	**0.326**	0.242	0.614	6.000	7.317
#6	T7	Digital & collaborative learning	178	0.351	0.246	**0.603**	15.000	2.093
	T9	Democracy literacy	185	0.333	0.246	0.614	9.000	13.441
	T12	Schools & teachers	187	0.345	**0.290**	0.610	13.000	3.986

Topic modeling

Dirichlet-Multinomial regression (DMR) (Mimno and McCallum, 2012)

Topic modeling parameters

Nr of topics: from 5 to 40

Nr. of topics

Label

15 manually assigned multi word labels representing literacy domains

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Interdisciplinarity
Domain (corpus): Literacy

Problem statement

This study explores the topic-based interdisciplinarity in the research domain of literacy. Topic modeling uncovered 15 dominant research topics in the literacy field, as well as their up-and-down trends from 2000 to 2021.
For each topic, keywords were then replaced with disciplines (i.e. the labels), and interdisciplinarity was measured using four indices: variety, balance, disparity, and diversity. Finally, the interdisciplinarity of each topic, connectivity between topics, and topic trends were comprehensively analyzed on the keyword co-occurrence network.

Corpus

Origin: Scopus
Nr. of documents: 346,387
Details:

- In order to form a corpus that encompasses various disciplines, literacy is chosen as the target research domain.
    - Using the Scopus API, abstracts and several metadata of documents containing the word 'literacy' in titles were collected without any subject area limitation.
    - The domain is comprised of diverse topics with terms such as "digital literacy, information literacy, data literacy, media literacy, early literacy, health literacy, and physical literacy"
- Abstracts of 346,387 articles published in 296 disciplines from 1917 to 2021.
- The search scope was limited to article and conference papers as document type and English as the language.

Document

Title, keywords and abstract of a Scopus paper in the literacy domain

Pre-processing

To collect keywords from abstracts and titles using a bottom-up approach, we adopted a deep keyword generation model.
This model represents the semantic meaning of a given text by using an RNN Encoder-Decoder model.
By incorporating a copying mechanism, RNN can predict out-of-vocabulary words, making it possible to generate topically relevant keywords that are absent from the text.
This method produces a list of candidate keywords from the text and divides them into multiple text chunks using n-grams or noun phrases.
Then, keywords are ranked with machine-learning features such as TF-IDF and PageRank, and the most meaningful keywords are selected.

```
@article{kim_2022_developing_a_topic_driven_method_for_interdisciplinarity_anal
ysis,
  abstract = {This study explores the topic-based interdisciplinarity in the
research domain of literacy. A text corpus of keywords was generated through a
deep keyword generation model from abstracts of 346,387 articles published in
296 disciplines from 1917 to 2021. Dirichlet-Multinomial Regression topic
modeling, interdisciplinarity indices, and network analysis were employed to
analyze the collected corpus. Topic modeling uncovered 15 dominant research
```

topics in the literacy field, as well as their up-and-down trends from 2000 to 2021. For each topic, keywords were then replaced with disciplines, and interdisciplinarity was measured using four indices: variety, balance, disparity, and diversity. Finally, the interdisciplinarity of each topic, connectivity between topics, and topic trends were comprehensively analyzed on the keyword co-occurrence network. Our methodology reaches beyond connectivity limited to a few disciplines and provides insight into the direction of collaboration between disciplines centered on a research domain. Moreover, the study's deep keyword generation model has methodological implications for forming a corpus spanning numerous disciplines as a bottom-up approach.},
  author = {Hyeyoung Kim and Hyelin Park and Min Song},
  date-added = {2023-03-14 19:34:49 +0100},
  date-modified = {2023-03-14 19:34:49 +0100},
  doi = {https://doi.org/10.1016/j.joi.2022.101255},
  issn = {1751-1577},
  journal = {Journal of Informetrics},
  keywords = {Disciplinarity, Interdisciplinary cooperation, Topic diversity, Keyword generation, DMR topic modeling, Deep learning},
  number = {2},
  pages = {101255},
  title = {Developing a topic-driven method for interdisciplinarity analysis},
  url = {https://www.sciencedirect.com/science/article/pii/S1751157722000074},
  volume = {16},
  year = {2022}}

#Thesis/Papers/Initial