kim_2022_exploring_scientific_trajectories_of_a_large_scale_dataset_using_topic_integrated_path_extraction

Year

2022

Author(s)

Erin H.J. Kim and Yoo Kyung Jeong and YongHwan Kim and Min Song

Title

Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction

Venue

Journal of Informetrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

"Biology and information science experts reviewed the LDA topic modeling results and labeled topic names in light of the extent of healthcare informatics subfields"

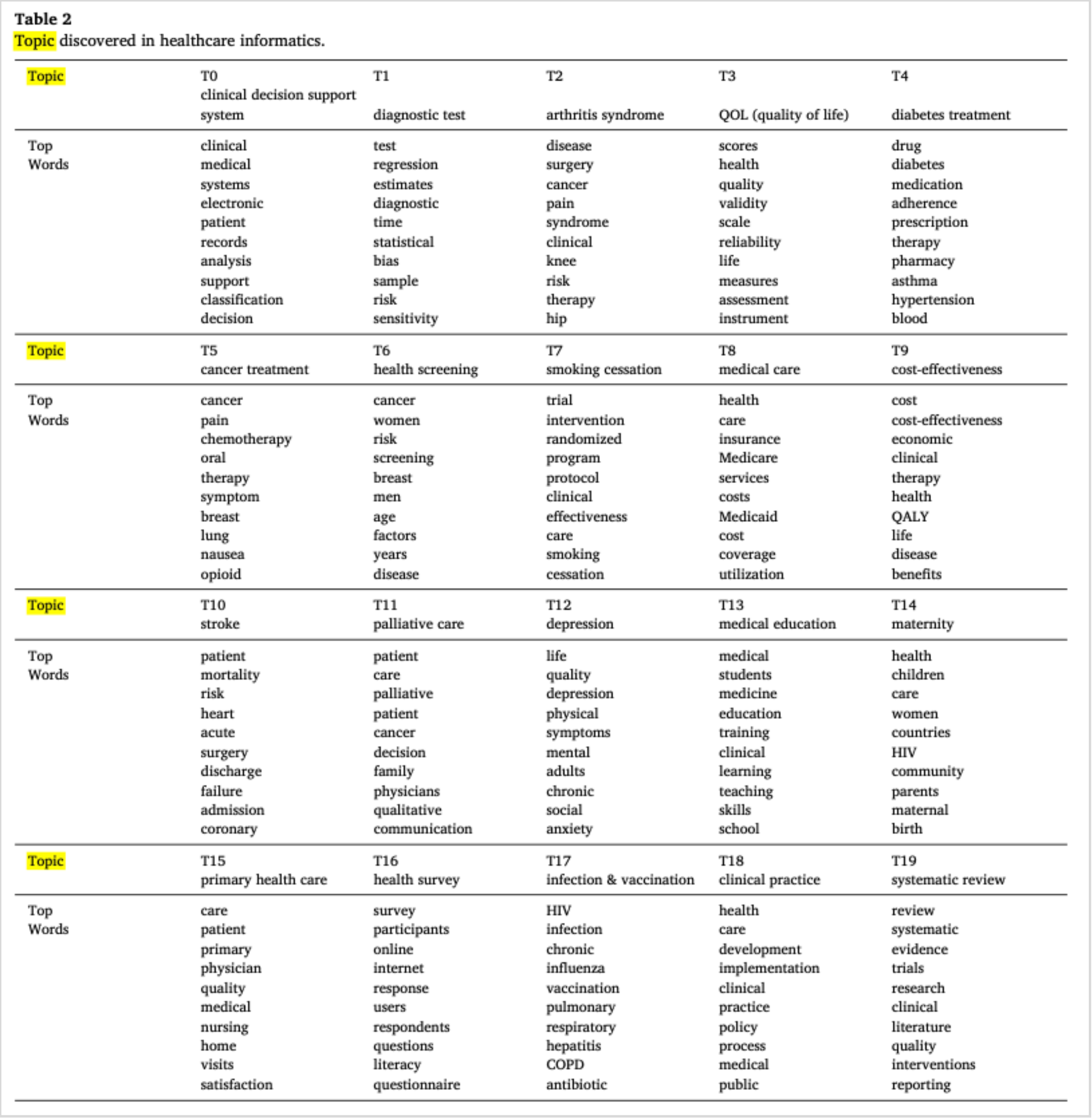

Table 2
Topic discovered in healthcare informatics.

Topic	T0	T1	T2	T3	T4
	clinical decision support system	diagnostic test	arthritis syndrome	QOL (quality of life)	diabetes treatment
Top Words	clinical	test	disease	scores	drug
	medical	regression	surgery	health	diabetes
	systems	estimates	cancer	quality	medication
	electronic	diagnostic	pain	validity	adherence
	patient	time	syndrome	scale	prescription
	records	statistical	clinical	reliability	therapy
	analysis	bias	knee	life	pharmacy
	support	sample	risk	measures	asthma
	classification	risk	therapy	assessment	hypertension
	decision	sensitivity	hip	instrument	blood
Topic	T5	T6	T7	T8	T9
	cancer treatment	health screening	smoking cessation	medical care	cost-effectiveness
Top Words	cancer	cancer	trial	health	cost
	pain	women	intervention	care	cost-effectiveness
	chemotherapy	risk	randomized	insurance	economic
	oral	screening	program	Medicare	clinical
	therapy	breast	protocol	services	therapy
	symptom	men	clinical	costs	health
	breast	age	effectiveness	Medicaid	QALY
	lung	factors	care	cost	life
	nausea	years	smoking	coverage	disease
	opioid	disease	cessation	utilization	benefits
Topic	T10	T11	T12	T13	T14
	stroke	palliative care	depression	medical education	maternity
Top Words	patient	patient	life	medical	health
	mortality	care	quality	students	children
	risk	palliative	depression	medicine	care
	heart	patient	physical	education	women
	acute	cancer	symptoms	training	countries
	surgery	decision	mental	clinical	HIV
	discharge	family	adults	learning	community
	failure	physicians	chronic	teaching	parents
	admission	qualitative	social	skills	maternal
	coronary	communication	anxiety	school	birth
Topic	T15	T16	T17	T18	T19
	primary health care	health survey	infection & vaccination	clinical practice	systematic review
Top Words	care	survey	HIV	health	review
	patient	participants	infection	care	systematic
	primary	online	chronic	development	evidence
	physician	internet	influenza	implementation	trials
	quality	response	vaccination	clinical	research
	medical	users	pulmonary	practice	clinical
	nursing	respondents	respiratory	policy	literature
	home	questions	hepatitis	process	quality
	visits	literacy	COPD	medical	interventions
	satisfaction	questionnaire	antibiotic	public	reporting

Motivation

Identifying healthcare informatics subfields represented by each label

Topic modeling

LDA

Topic modeling parameters

Nr of topics (k): 10 to 30

Nr. of topics

20

Label

Single or multi-word label identifying healthcare informatics subfields

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Citation analysis
Dataset: Healthcare informatics

Problem statement

Main path analysis (MPA) is the most widely accepted approach to tracing knowledge transfer in a research field.
In this study, we extracted multiple longest paths from the multidisciplinary academic field's citation network and integrating topic modeling to the extracted paths. We consider three main aspects of trajectory analysis when analyzing the represented documents through the extracted paths: emergence, authority, and topic dynamics.
For topic integration into multiple paths, we employ latent Dirichlet allocation (LDA) by

utilizing the topic-document matrix that LDA derives to select an article's topic from the citation network, where each article is labeled with the topic that is assigned with the highest topical probability for that article.

Corpus

Origin: PubMed
Nr. of documents: 274,297 papers and 595,548 citing-cited pairs
Details:

- Healthcare informatics, seed articles from top 30 journals in the healthcare informatics field based on JCR reports
- Publication years from 1970 to 2017
- 89,369 seed papers are collected, from these additional citing papers are collected

Document

Paper in the healthcare informatics domain with journal title, authors, title, abstract, and publication year

Pre-processing

If the PubMed ID (PMID) of a cited paper was equal to the PMID of a citing paper (i.e., if paper A cited paper A' and paper A was identical to A'), that pair was removed. Additionally papers with self-citations and citation errors are discarded

---

```
@article{kim_2022_exploring_scientific_trajectories_of_a_large_scale_dataset_using_topic_integrated_path_extraction,
  abstract = {Main path analysis (MPA) is the most widely accepted approach to
tracing knowledge transfer in a research field. In this study, we extracted
multiple longest paths from the multidisciplinary academic field's citation
network and integrating topic modeling to the extracted paths. We consider
three main aspects of trajectory analysis when analyzing the represented
documents through the extracted paths: emergence, authority, and topic
dynamics. For path extraction, we adopt the longest path algorithm that
consists of the following three steps: 1) topological sort, 2) edge relaxation,
and 3) multiple path extraction. For topic integration into multiple paths, we
employ latent Dirichlet allocation (LDA) by utilizing the topic-document matrix
```

```
that LDA derives to select an article's topic from the citation network, where
each article is labeled with the topic that is assigned with the highest
topical probability for that article. We conduct a series of experiments to
examine the results on a dataset from the field of healthcare informatics that
PubMed provides.},
  author = {Erin H.J. Kim and Yoo Kyung Jeong and YongHwan Kim and Min Song},
  date-added = {2023-03-15 19:41:40 +0100},
  date-modified = {2023-03-15 19:41:40 +0100},
  doi = {https://doi.org/10.1016/j.joi.2021.101242},
  issn = {1751-1577},
  journal = {Journal of Informetrics},
  keywords = {Citation analysis, Healthcare informatics, Longest path, Main
path analysis, Topic modeling},
  number = {1},
  pages = {101242},
  title = {Exploring scientific trajectories of a large-scale dataset using
topic-integrated path extraction},
  url = {https://www.sciencedirect.com/science/article/pii/S1751157721001139},
  volume = {16},
  year = {2022}}
```

#Thesis/Papers/Initial