truic_2021_tlatr_automatic_topic_labeling_using_automatic_domain_specific_term_recognition

Year

2021

Author(s)

Ciprian-Octavian Truică and Elena Simona Apostol

Title

TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition

Venue

IEEE Access

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Automatic (domain-specific) term recognition algorithm (ATR)

Topic labeling parameters

Extracted n-gram size: n ≥ 2

Label generation

Introduction

For determining the terms that are relevant to a topic, C-Value is used for extracting and scoring the domain-specific terms

The topic label is chosen as the domain-specific term with the highest C-Value score.

TLATR consists of two main steps.

Firstly, it incorporates the context dimension to determine a list of representative domain-specific candidate terms.

These candidate terms are extracted from the documents belonging to a topic using linguistic filters.

Secondly, we extract the labels for each topic from the list of candidate terms.

Our method is strictly unsupervised, i.e., there is no labeled corpus from which to build the model. Everything is learned dynamically, no model is built and the labels are highly dependent on the quality of the topics and the documents belonging to these topics.

Methodology

Our method, TLATR, uses automatic domain- specific term recognition to extract the candidate n-grams (*cng*) that are relevant to the documents *D* belonging to a topic *z*. The candidate n-grams are then scored to obtain the list of relevant domain-specific terms (*dst*). The term with the highest score is chosen from domain-specific terms (*dst*) list as label *l* for a topic *z*.

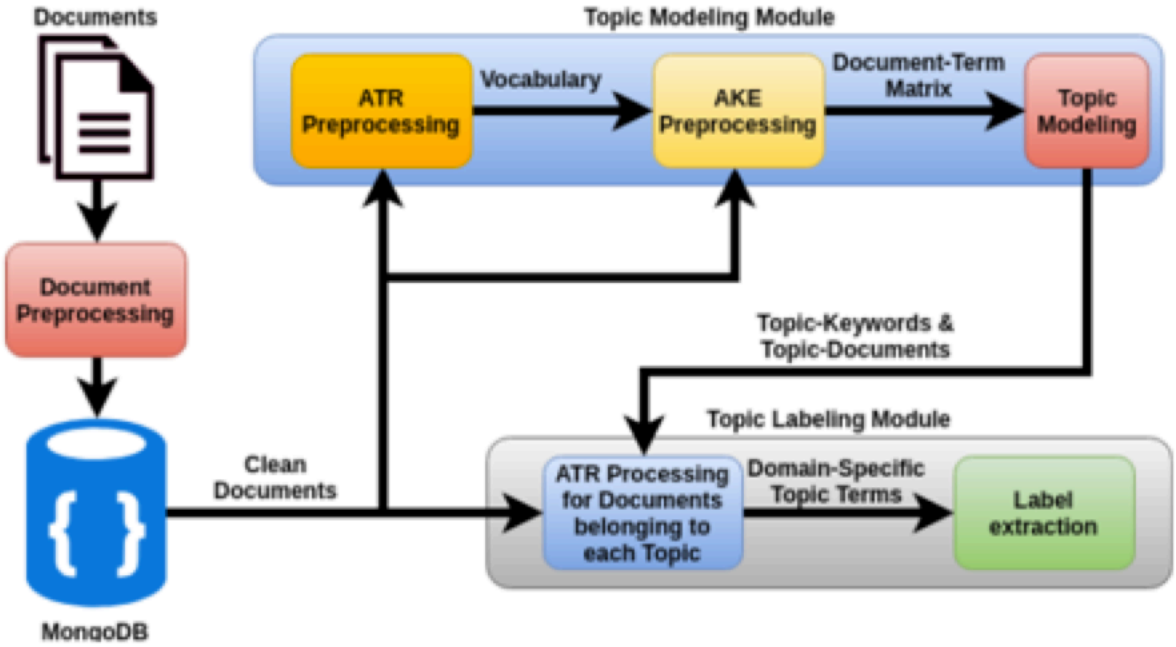

FIGURE 1. Logical schema for the proposed method.

Topic modeling module

1. *ATR Preprocessing*: during this stage we extract the list of domain-specific terms (*dst*) using Automatic Term Recognition (ATR) and construct the vocabulary;
    - One method used to extract domain-specific multi-word terms is C-Value

$$\text{C-value(a)} = \begin{cases} \log_2 |a| \cdot f(a) \text{ where a is not nested} \\ \log_2 |a| \cdot (f(a) - \frac{1}{P(T_a)} \\ \cdot \sum_{b \in T_a} f(b)) \text{ otherwise} \end{cases}$$

    a is a candidate string, |*a*| is the number of terms in *a*, *f* (*a*) is the frequency of *a* in the corpus, T_a is the list of candidate terms (n-grams) that contain *a* and *P*(T_a) is the number of these candidate terms (n-grams)
    - The linguistic filters used for extracting the candidate string are user-defined and usually, they contain Nouns, Adjectives, and Noun Prepositions
2. *AKE Preprocessing*: this second stage implies building the document-term matrix employing Automatic Keyword Extraction (AKE);
    - For constructing the document-term matrix after the corpus vocabulary is determined, we use *TF-IDF* and *Okapi BM25*.
3. *Topic Extraction*: in the last stage we apply topic mod-eling algorithms to extract topics.
    - `see topic modeling`

Topic labeling module

This module consists of two components.

The first component applies ATR on each group of documents belonging to each topic, similar to the one discussed in before, with the difference that here we use the original version of C-Value where only n-grams with $n \geq 2$ are extracted.

The second component uses the output of the ATR pre-processing component to extract and select from the domain-specific list the term with the highest score as topic label. Thus, ATR is used for automatic label detection by extracting n-grams using linguistic patterns and the topic labels are determined from the list of the domain-specific terms extracted from the documents belonging to the same topic.

Details of the algorithm

The proposed method is implemented through Algorithm 1.

Algorithm 1 TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition

Require: a topic-documents set $TD = \{(z_i, D_i) | D_i \text{ is the set of documents for topic } z_i\}$ and the list of linguistic patterns LP

Ensure: a topic-label set $TL = \{(z_i, l_i) | l_i \text{ is the label for topic } z_i\}$

```
TL = ∅
for each (z_i, D_i) ∈ TD do in parallel
    cng = ∅
    dst = ∅
    for d ∈ D_i do
        d = expandContractions(d)
        for sentence ∈ d do
            lemmaPOS = ∅
            tokens = wordTokenize(sentence)
            tokensPOS = posTokenize(tokens)
            for token, pos ∈ tokensPOS do
                lemma = lemmatization(token, pos)
                lemmaPOS = lemmaPOS ∪ {(lemma, pos)}
            cng = cng ∪ extractNGrams(lemmasPOS, LP)
    for ngram ∈ cng do in parallel
        dst = dst ∪ {(ngram, C-Value(ngram))}
    dst = sort(dst, key=C-Value)
    l_i = dst[0]
    TL = TL ∪ {(z_i, l_i)}
return TL
```

The input is a topic-documents set (*TD*) that contains the list of documents assigned to each topic and a list of linguistic patterns (*LP*) needed to extract the candidate n-grams (*cng*) before applying *C-Value* and determining the domain-specific terms (*dst*).

The output is a topic-label set (*TL*) that contains the label *l_i* for a topic *z_i*.

The linguistic patterns (*LP*) has a dual-purpose:

1. adding context to the topic labeling through the use of the part of speech patterns, and
2. filtering the words taken into account for determining the domain-specific terms (*dst*)

During the iteration of the topic-documents set (*TD*), the candidate n-grams (*cng*) for each topic are

extracted.

The candidate n-grams (*cng*) list contains all the candidate n-grams for all the documents labeled with the same topic. To extract the candidate n-grams (*cng*), each document assigned to a topic is processed as follows (Lines 5 to 14):

1. the contractions are expanded (Line 6);
2. the document is split into sentences (Line 7);
3. for each sentence the tokens and their corresponding part of speech (*pos*) and lemmas are extracted (Lines 8 to 13);
4. the candidate n-grams (*cng*) is constructed using the list of linguistic patterns (*LP*) and the lemma and part of speech of each term (Line 14).

During the iteration of the candidate n-grams (*cng*) list, C-Value is used to extract and score the domain-specific terms (*dst*) are then sorted in descending order by their C-Value score (Line 17). The topic label is determined (Line 18) and the topic-label set (*TL*) is updated (Line 19).
When labels are determined for all the topics, the algorithm finishes and returns the topic-label set (*TL*) (Line 20).

Results

ART DATASET

TABLE 4. ART NMF TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	data mining method set approach algorithm paper pattern classification large	4.75	0.55	0.80	3.00
2	database system	query database relational performance xml object processing optimization language data	4.89	0.54	0.80	3.00
3	real time	image surface method object model motion point shape scene camera	7.37	0.71	0.65	2.43
4	low bound	problem algorithm time bound graph approximation log number case polynomial	7.29	0.81	0.55	2.65
5	research paper	system information application research technology user management paper data visualization	5.74	0.57	0.87	2.88

TABLE 5. Compassion TLART vs NETL on ART NMF TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	data set	methodology	**0.80**	0.78	0.65
2	database system	relational database	relational database	0.80	**0.84**	**0.84**
3	real time	camera	camera	**0.65**	**0.65**	**0.65**
4	low bound	polynomial	polynomial	0.55	**0.66**	**0.66**
5	research paper	information management	information management	**0.87**	0.84	0.84

TABLE 6. ART LDA TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	system data research database information paper knowledge application technology medical	5.73	0.61	0.80	3.00
2	database system	data query database performance algorithm paper system result approach technique	5.08	0.50	0.85	3.00
3	real time	image method model surface motion visualization object shape reconstruction approach	7.23	0.70	0.63	2.36
4	low bound	bound algorithm problem time complexity polynomial low number random case	7.25	0.84	0.66	2.67
5	approximation algorithm	graph algorithm problem approximation time edge vertex minimum tree log	5.08	0.57	0.87	2.76

TABLE 7. Compassion TLATR vs NETL on LDA TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	information system	information system	0.80	**0.85**	**0.85**
2	database system	database	database	**0.85**	0.69	0.69
3	real time	visualization computer graphics	approximation	**0.63**	0.62	0.40
4	low bound	polynomial	polynomial	**0.66**	0.57	0.57
5	approximation algorithm	minimum spanning tree	optimization problem	**0.87**	0.69	0.66

TABLE 8. ART NMF Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data mining	data mining visualization rule discovery knowledge association pattern approach technique	4.56	0.53	0.77	3.00
2	database system	database system query object management relational language xml processing performance	4.03	0.47	0.83	3.00
3	real time	image model segmentation motion reconstruction method surface registration object shape	7.47	0.70	0.61	2.37
4	low bound	algorithm problem time graph approximation result number bound tree polynomial	7.50	0.81	0.55	2.36
5	research paper	research information medical technology paper health application care clinical patient	5.60	0.64	0.84	2.85

TABLE 9. Compassion TLART vs NETL on NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data mining	data mining	information visualization	**0.77**	**0.77**	0.74
2	database system	relational database	relational database	**0.83**	**0.83**	**0.83**
3	real time	image registration	image registration	0.61	**0.71**	**0.71**
4	low bound	polynomial	polynomial	0.55	**0.65**	**0.65**
5	research paper	medical research	clinical research	0.84	**0.91**	0.85

TABLE 10. ART LDA Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	data set	data image method algorithm model approach feature object pattern classification	5.25	0.56	0.74	3.00
2	database system	database data system query management application web information object mining	4.38	0.46	0.88	3.00
3	black box	abstract reconstruction tree visualization preliminary dynamic pet interactive computation data	10.35	0.99	0.77	2.54
4	low bound	algorithm problem graph time bound approximation point polynomial number linear	7.50	0.81	0.58	2.57
5	research paper	medical image research system clinical patient paper neural information registration	6.34	0.72	0.81	2.91

TABLE 11. Compassion TLART vs NETL on LDA Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	data set	data model	data model	0.74	**0.81**	**0.81**
2	database system	web application	web application	**0.88**	0.80	0.80
3	black box	data visualization	data visualization	0.77	**0.79**	**0.79**
4	low bound	polynomial	polynomial	0.58	**0.69**	**0.69**
5	research paper	clinical research	clinical research	**0.81**	**0.81**	**0.81**

ARX DATASET

TABLE 12. ARX NMF TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	black hole horizon gravitational gravity entropy mass spacetime kerr binary	5.97	1.00	0.85	3.00
2	cross section	collision energy production section proton cross momentum transverse hadron tev	7.86	1.00	0.77	2.76
3	dark matter	higgs mass boson standard neutrino decay lhc model search gev	7.66	0.95	0.67	3.00
4	dark matter	cosmological universe matter dark inflation cosmic scale cosmology background energy	6.77	0.81	0.75	3.00
5	differential equation	equation solution differential nonlinear wave condition method problem function schr	5.78	0.62	0.85	2.96
6	field theory	theory field gauge symmetry gravity scalar action term conformal dimension	4.34	0.54	0.83	2.91
7	form factor	quark qcd meson mass pi chiral lattice decay bar heavy	8.81	0.93	0.77	2.79
8	low bound	algorithm problem graph optimization complexity number optimal time bound method	7.17	0.69	0.56	2.23
9	magnetic field	magnetic spin electron field effect interaction energy surface material electronic	6.13	0.72	0.81	2.93
10	mathbb r	space group algebra class manifold operator mathbb representation theorem case	6.77	0.67	0.48	2.39
11	monte carlo	distribution random method function probability model estimation parameter estimator data	10.12	1.00	0.45	2.31
12	neural network	network neural task image art feature deep learning performance convolutional	6.31	0.76	0.81	3.00
13	numerical simulation	dynamic time system simulation model flow particle equilibrium process dynamical	7.53	0.72	0.76	2.86
14	phase transition	phase transition temperature critical order lattice point diagram finite behavior	5.11	0.65	0.79	2.72
15	quantum system	quantum state system entanglement classical qubit mechanic information measurement entropy	3.26	0.34	0.91	2.97
16	single photon	optical frequency photon laser mode light high wave beam cavity	6.16	0.62	0.73	2.75
17	social network	data analysis research paper approach application recent information development tool	6.23	0.57	0.72	2.31
18	upper bound	channel rate capacity user scheme network transmission communication receiver information	8.23	0.81	0.38	2.23
19	x ray	star galaxy stellar mass observation emission ray formation line gas	7.98	0.89	0.51	2.43

TABLE 13. Compassion TLART vs NETL on ARX NMF TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	gravity	gravity	**0.85**	0.75	0.75
2	cross section	kinetic energy	pair production	**0.77**	0.67	0.47
3	dark matter	standard model	standard model	**0.67**	0.44	0.44
4	dark matter	cosmic microwave background	cosmology	**0.75**	0.71	0.68
5	differential equation	differential equation	differentsal equation	**0.85**	**0.85**	**0.85**
6	field theory	scalar field	scalar field	**0.83**	0.71	0.71
7	form factor	quark	quark	**0.77**	0.69	0.69
8	low bound	time complexity	time complexity	0.56	**0.77**	**0.77**
9	magnetic field	magnetic field	electron	**0.81**	**0.81**	0.69
10	mathbb r	linear algebra	representation theory	0.48	0.70	**0.72**
11	monte carlo	probability distribution	probability distribution	0.45	**0.81**	**0.81**
12	neural network	artificial neural network	artificial neural network	**0.81**	0.79	0.79
13	numerical simulation	dynamical system	dynamical system	0.76	**0.82**	**0.82**
14	phase transition	phase diagram	phase diagram	**0.79**	0.78	0.78
15	quantum system	quantum entanglement	quantum entanglement	**0.91**	0.78	0.78
16	single photon	photon	photon	**0.73**	0.66	0.66
17	social network	information	information	**0.72**	0.69	0.69
18	upper bound	computer network	telecommunication	0.38	**0.75**	0.54
19	x ray	star formation	star formation	0.51	**0.73**	**0.73**

TABLE 14. ARX LDA TF-IDF.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	star galaxy ray mass observation emission stellar source high line	9.42	1.00	0.76	3.00
2	black hole	theory field gauge black symmetry hole gravity quantum scalar equation	8.18	1.00	0.81	3.00
3	constacyclic code	phys rev lf lett al textbf paper fr uhlenbeck wavelet	10.41	0.92	0.76	2.21
4	cross section	quark qcd collision heavy calculation nuclear energy momentum nucleon meson	9.01	1.00	0.75	2.79
5	dark matter	dark cosmological universe matter inflation model scale planck cosmology cosmic	6.81	0.81	0.78	3.00
6	differential equation	equation function operator solution matrix problem graph eigenvalue result polynomial	7.18	0.67	0.83	2.93
7	functional theory	molecule molecular hydrogen van calculation ab li atom imito functional	7.89	0.80	0.81	2.71
8	low bound	graph network node model complexity logic population problem algorithm tree	8.24	0.75	0.51	2.21
9	low bound	channel information code capacity communication rate bound scheme bit error	6.82	0.68	0.61	2.27
10	magnetic field	system dynamic equation equilibrium particle numerical phase model time energy	7.34	0.75	0.76	2.71
11	magnetic field	spin magnetic temperature phase topological electron band state transition lattice	6.13	0.67	0.74	3.00
12	moduli space	group algebra manifold space lie class algebraic mathbb invariant representation	7.47	0.69	0.72	2.77
13	molecular dynamic	cell surface protein material elastic molecular concentration force simulation polymer	7.11	0.62	0.75	2.88
14	monte carlo	method algorithm problem estimation distribution data error model approach function	10.56	1.00	0.64	2.67
15	neural network	network task image neural art method data learning feature datasets	6.54	0.74	0.80	3.00
16	quantum mechanic	physic development research recent question theory understanding year mechanic concept	6.89	0.71	0.78	3.00
17	single photon	quantum optical photon frequency laser atom state light system mode	6.22	0.60	0.75	2.83
18	social network	network user paper performance strategy market data price system time	5.71	0.50	0.69	2.79
19	standard model	decay neutrino mass gev higgs lhc tev detector boson standard	6.81	0.79	0.69	2.43

TABLE 15. Compassion TLART vs NETL on ARX LDA TF-IDF.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	star	stellar evolution	**0.76**	0.55	0.68
2	black hole	quantum field theory	quantum field theory	0.81	**0.84**	**0.84**
3	constacyclic code	mathematical physics	mathematical physics	**0.76**	0.31	0.31
4	cross section	nucleon	nucleon	**0.75**	0.66	0.66
5	dark matter	cosmology	cosmology	**0.78**	0.73	0.73
6	differential equation	schrodinger equation	rotation matrix	**0.83**	0.54	0.66
7	functional theory	molecule	molecule	**0.81**	0.75	0.75
8	low bound	time complexity	computational complexity theory	0.51	0.69	**0.71**
9	low bound	bit rate	communications protocol	0.61	**0.72**	0.63
10	magnetic field	partial differential equation	euler equations fluid dynamics	**0.76**	0.75	0.72
11	magnetic field	electron	electron	**0.74**	0.68	0.68
12	moduli space	lie algebra	lie algebra	0.72	**0.77**	**0.77**
13	molecular dynamic	polymer	polymer	**0.75**	0.60	0.60
14	monte carlo	normal distribution	normal distribution	0.64	**0.69**	**0.69**
15	neural network	artificial neural network	database	0.80	**0.81**	0.60
16	quantum mechanic	research	research	**0.78**	0.74	0.74
17	single photon	laser	laser	**0.75**	0.69	0.69
18	social network	market data	market data	0.69	**0.85**	**0.85**
19	standard model	neutrino	neutrino	0.69	**0.71**	**0.71**

TABLE 16. ARX NMF Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	black hole horizon gravitational gravity entropy spacetime ad kerr schwarzschild	5.95	1.00	0.82	3.00
2	correlation function	function operator term formula matrix expansion expression integral form coefficient	5.57	0.54	0.82	2.87
3	cross section	collision production quark energy section proton cross momentum heavy hadron	8.33	1.00	0.78	2.78
4	dark matter	mass higgs standard boson decay neutrino model lhc search gev	7.65	0.95	0.67	3.00
5	dark matter	cosmological universe inflation matter scalar dark gravity scale model energy	6.70	0.76	0.77	3.00
6	differential equation	equation solution differential nonlinear condition schr wave existence odinger problem	5.77	0.63	0.86	3.00
7	field theory	theory field gauge symmetry action gravity conformal effective loop dimension	4.37	0.54	0.83	3.00
8	lie algebra	group algebra representation lie finite algebraic module category product subgroup	5.56	0.57	0.75	2.76
9	low bound	graph number vertex edge bound set degree random tree polynomial	7.43	0.75	0.65	2.23
10	magnetic field	optical frequency photon laser mode light wave high beam cavity	6.98	0.66	0.75	2.84
11	magnetic field	phase transition temperature magnetic spin lattice density interaction state critical	6.45	0.72	0.75	2.87
12	moduli space	space manifold dimensional surface dimension condition geometry theorem point curvature	6.49	0.64	0.74	2.67
13	monte carlo	distribution random probability model process estimator parameter estimation statistical asymptotic	10.23	1.00	0.45	2.63
14	neural network	network neural task data feature art learning image model deep	6.60	0.71	0.80	3.00
15	numerical simulation	dynamic system time particle simulation equilibrium dynamical numerical evolution model	7.51	0.71	0.84	2.97
16	optimization problem	algorithm problem method optimization paper computational approach technique performance numerical	5.29	0.52	0.85	2.95
17	quantum system	quantum state entanglement classical system mechanic qubit information measurement entropy	9.35	0.96	0.91	3.00
18	upper bound	channel rate capacity scheme information communication user transmission receiver code	8.08	0.80	0.38	2.25
19	x ray	star galaxy observation ray mass stellar emission formation source high	7.96	0.87	0.52	2.49

TABLE 17. Compassion TLART vs NETL on ARX NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	black hole	gravitational field	**0.82**	**0.82**	0.73
2	correlation function	wave function	wave function	**0.82**	0.75	0.75
3	cross section	pair production	pair production	**0.78**	0.54	0.54
4	dark matter	standard model	standard model	**0.67**	0.44	0.44
5	dark matter	dark matter	cosmological constant	**0.77**	**0.77**	0.75
6	differential equation	differential equation	differential equation	**0.86**	**0.86**	**0.86**
7	field theory	quantum field theory	lagrangian field theory	**0.83**	0.79	0.64
8	lie algebra	lie algebra	algebraic group	0.75	0.75	**0.81**
9	low bound	chromatic polynomial	chromatic polynomial	**0.65**	0.57	0.57
10	magnetic field	photon	photon	**0.75**	0.66	0.66
11	magnetic field	magnetic field	magnetization	**0.73**	**0.73**	0.46
12	moduli space	curvature	curvature	**0.74**	0.64	0.64
13	monte carlo	probability distribution	probability distribution	0.45	**0.84**	**0.84**
14	neural network	artificial neural network	artificial neural network	**0.80**	0.75	0.75
15	numerical simulation	physical system	euler equations fluid dynamics	**0.84**	0.69	0.68
16	optimization problem	optimization problem	optimization problem	**0.85**	**0.85**	**0.85**
17	quantum system	quantum entanglement	quantum entanglement	**0.91**	0.86	0.86
18	upper bound	information	information	0.38	**0.63**	**0.63**
19	x ray	star formation	star formation	0.52	**0.73**	**0.73**

TABLE 18. ARX LDA Okapi BM25.

# Topic	Label	Topic keywords	PMI	NPMI	$cos(\theta)$	AvgR
1	black hole	star mass gravitational binary hole black disk radius galaxy rotation	8.07	1.00	0.81	3.00
2	black hole	gravity field scalar black hole cosmological theory einstein spacetime gravitational	7.83	1.00	0.83	3.00
3	cross section	collision quark production heavy decay mass section lhc gev energy	8.71	1.00	0.79	2.78
4	dark matter	dark neutrino matter cosmic background experiment data ray mass energy	7.09	0.92	0.75	3.00
5	differential equation	equation solution system time function numerical dynamic nonlinear differential condition	6.76	0.66	0.88	3.00
6	field theory	theory gauge symmetry loop field fermion model lattice operator chiral	5.07	0.57	0.82	3.00
7	graph g	graph entropy edge vertex entanglement random number tree state node	4.44	0.42	0.58	3.00
8	lie algebra	algebra group manifold space algebraic representation lie invariant construction class	6.11	0.58	0.76	2.87
9	low bound	bound proof function problem polynomial set mathbb result theorem paper	7.53	0.73	0.65	2.24
10	macwilliams identity	math triangle cusp adequate cc wedge cd versa angle vice	10.73	0.98	0.68	2.67
11	magnetic field	magnetic electron spin field temperature optical phase effect state transition	6.34	0.67	0.80	3.00
12	nash equilibrium	optimal strategy problem game market cost paper price time agent	9.49	0.84	0.79	2.77
13	neural network	method algorithm data problem approach performance estimation model paper task	7.88	0.82	0.73	3.00
14	numerical simulation	flow simulation force fluid dynamic cell surface model pressure particle	7.21	0.67	0.71	2.89
15	quantum system	quantum state carlo monte system measurement coherent photon qubit single	3.61	0.37	0.85	3.00
16	real time	development research physic science technology application year progress design challenge	7.04	0.59	0.60	2.67
17	social network	network population human individual social model data pattern activity dynamic	5.29	0.50	0.84	3.00
18	upper bound	channel code capacity communication rate transmission bit scheme error information	8.03	0.80	0.39	2.27
19	x ray	emission observation star line ray solar high source galaxy telescope	8.09	0.90	0.52	2.48

TABLE 19. Compassion TLART vs NETL on LDA NMF Okapi BM25.

	Labels			$cos(\theta)$		
# Topic	TLATR	SNETL	UNETL	TLATR	SNETL	UNETL
1	black hole	binary star	binary star	**0.81**	0.70	0.70
2	black hole	gravitational field	gravitational field	**0.83**	0.80	0.80
3	cross section	radioactive decay	radioactive decay	**0.79**	0.66	0.66
4	dark matter	cosmic microwave background	cosmic microwave background	0.75	**0.77**	**0.77**
5	differential equation	differential equation	differential equation	**0.88**	**0.88**	**0.88**
6	field theory	quantum field theory	supersymmetry	**0.82**	0.75	0.46
7	graph g	vertex cover	vertex cover	**0.58**	0.47	0.47
8	lie algebra	lie algebra	lie algebra	**0.76**	**0.76**	**0.76**
9	low bound	polynomial	polynomial	**0.65**	0.61	0.61
10	macwilliams identity	right triangle	right triangle	**0.68**	0.70	0.70
11	magnetic field	magnetic field	electromagnetic field	**0.80**	**0.80**	0.72
12	nash equilibrium	optimization problem	optimization problem	**0.79**	0.66	0.66
13	neural network	mathematical model	optimization problem	0.73	**0.74**	**0.74**
14	numerical simulation	fluid dynamics	fluid dynamics	0.71	**0.82**	**0.82**
15	quantum system	quantum system	quantum entanglement	**0.85**	**0.85**	0.71
16	real time	design research	design research	0.60	**0.84**	**0.84**
17	social network	social network	social network	**0.84**	**0.84**	**0.84**
18	upper bound	channel state information	channel state information	0.39	**0.79**	**0.79**
19	x ray	hubble space telescope	hubble space telescope	0.52	**0.66**	**0.66**

Motivation

\

Topic modeling

LDA, NMF

Topic modeling parameters

Nr of topics: 19 + 5

Nr. of topics

19 + 5

Label

N >= 2 n-gram extracted from document

Label selection

\

Label quality evaluation

The topic label evaluation is done both automatically and using human annotators.

Automatic evaluation

As evaluation methods for the Topic labeling module, we utilize Pointwise Mutual Information (PMI), Normalized Point- wise Mutual Information (NPMI), and cosine similarity between the average word vectors of a topic label and of its keywords.

In the context of topic labeling, PMI determines the relevance of a label to a topic, treating each label as a collocation (i.e. a sequence of words or terms that co-occur more often than would be expected

by chance). By computing the PMI for a label using all the collocations that appear in the documents belonging to a topic, it can be determined if a label is indeed representative for a topic.
The PMI of a pair of outcomes *x* and *y* belonging to discrete random variables *X* and *Y* quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence.

In the context of topic labeling, NPMI normalizes the score given by PMI for a label. So the closest the score is to the upper bound, the more relevant is that label for the set of documents belonging to that topic.

To compute PMI and NPMI, we apply the following steps:

1. determine the documents that belong to each topic (a document is assigned to a single topic);
2. extract the n-grams from the subset of documents belonging to a topic;
3. compute the score using the label and the n-grams.

We use word vectors to estimate the semantic similarity between the labels and the topics. Both the topic labels and the keywords for the topic are transformed into word vectors using word embedding, representing each topic label (*TL*) and topic keywords (*TK*) as an average of their word vectors.

Vectors can then be compared using their cosine similarity

Human evaluation

To evaluate the quality of TLATR, we compute the Average Rating (AvgR) score by asking human annotator to rate a topic on a 4-point Likert Scale as follows:

1. A perfect label that describes the topic is graded with 3 (Very good label);
2. A label that is related to the topic but does not completely capture the topic is graded with 2 (Reasonable label);
3. A label that is semantically related to the topic but does not describe the topic very well is graded with 1 (Semantically related); but would not make a good topic label;
4. A label that is unrelated to the topic is graded with 0 (Inappropriate label).

Equation (7) presents the formula for computing the Average Rating, where n is the number of annotator and $s_j(z_i, l_i)$ $(j = \overline{1, n})$ is the grade given by each annotator of how well the label l_i describes topic z_i.

$$AvgR(z_i, l_i) = \frac{\sum_{j=1}^{n} s_j(z_i, l_i)}{n} \quad (7)$$

The annotators received the topic label, the topic's top-10 relevant keywords, and a grading scale from 0 to 3.
They were asked to grade, using this scale, how well the label describes the corresponding topic

We compare the topic labels obtained by our method with the labels obtained by NETL [5]. NETL offers two distinct topic labeling models: i) a supervised NETL (SNETL) model, and ii) an unsupervised NETL (UNETL) model. We used the cosine similarity for this comparison, as the task would be too tenuous for human annotators.

Assessors

30 graduate and undergraduate students of Computer Science

Domain

Paper: Topic labeling
Dataset: Scientific literature

Problem statement

In this paper, we present a new topic labeling method that uses automatic term recognition to discover and assign relevant labels for each topic, i.e., TLATR (Topic Labeling using Automatic Term Recognition). TLATR uses domain-specific multi-terms that appear in the set of documents belonging to a topic. The multi-term having the highest score as determined by the automatic term recognition algorithm is chosen as the label for that topic.

Corpus

ART

Origin: Aminer

Nr. of documents: 18464

Details:

- scientific articles abstracts

ARX

Origin: ArXiv

Nr. of documents: 166512

Details:

TABLE 2. Corpora labels.

Corpus	Labels	No. Documents
ART	database	2 270
	data mining	5 059
	medical	3 066
	theory	3 995
	visualization	4074
ARX	cs	21 999
	math	24 000
	physics	12 000
	physics_astro-ph	12 000
	physics_cond-mat	12 000
	physics_gr-qc	7 685
	physics_hep-ex	4 147
	physics_hep-lat	1 883
	physics_hep-ph	9 736
	physics_hep-th	9 572
	physics_math-ph	7 113
	physics_nlin	3 303
	physics_nucl-ex	2 045
	physics_nucl-th	4 255
	physics_physics	12 000
	physics_quant-ph	9 908
	q-bio	3 608
	q-fin	1 399
	stat	7 859

Document

Pre-processing

- text cleaning by removing HTML/XML tags and scripts
- language identification
- expanding contractions
- extracting the sentences
- applying the part of speech tagging and extract lemmas
- removing stop words and punctuation
- removing duplicate terms

```
@article{truic_2021_tlatr_automatic_topic_labeling_using_automatic_domain_specif
ic_term_recognition,
title={TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term
Recognition},
author={Ciprian-Octavian Truică and Elena Simona Apostol},
journal={IEEE Access},
year={2021},
volume={9},
pages={76624-76641}
}
```

#Thesis/Papers/FS