Measuring Tech Emergence Session

A “Contest”: Novel and Viable Indicators Using WoS Bibliographic Records

Measuring Tech Emergence Overview

- ***Why?*** Recognize value of understanding tech emergence and related concepts
  - study scientific evolution
  - individual science or technology or innovation indicators
  - composite indicators (e.g., dashboards) for science policy or technology management
  - competitive technical intelligence by identifying key players in particular frontier topics
  - and so on…

Measuring Tech Emergence Overview

- "Centers of gravity":
  - **IARPA FUSE** (U.S. Intelligence Advanced Research Projects Activity, Foresight & Understanding from Scientific Exposition) Program promoted development (2010-)
  - [SRI program] **NSF SciSIP** (EAGER) support: "Using the ORCID ID and Emergence Scoring to Study Frontier Researchers" (2016-2018)Underway with
  - **NSF SciSIP/NCSES** support: "Indicators of Technological Emergence" (April, 2018- March, 2021)

We acknowledge support from the US National Science Foundation (Award #1759960 – "Indicators of Technological Emergence") to Search Technology, Inc., and Georgia Tech. The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

A sampling of our experiences…

- Carley, S.F., Newman, N.C., Porter, A.L., and Garner, J. (2018). An indicator of technical emergence, *Scientometrics,* 115 (1), 35-49; http://link.springer.com/article/10.1007/s11192-018-2654-5.

- Porter, A.L., Garner, J., Carley, S.F., and Newman, N.C. (2018). Emergence scoring to identify frontier R&D topics and key players, *Technological Forecasting and Social Change*; https://doi.org/10.1016/j.techfore.2018.04.016.

- Shapira P, Kwon S, Youtie J. Tracking the Emergence of Synthetic Biology, Scientometrics, 2017, 112: 1439-1469. http://dx.doi.org/10.1007/s11192-017-2452-5.

Measuring Tech Emergence Contest
Can you devise a better way?

- To "reach out of our box," we set up a contest
  - treat words and phrases differently?
  - combine multiple WoS fields' content?
  - exploit other data attributes?
- Given a tech domain dataset (WoS records) for a 12 year period, **predict sub-topics to be heavily researched in the following 2 years**
- Thanks to Clarivate for XML Web of Science datasets!
  - ***Practice***: Neurodegenerative; Dye Sensitized Solar Cells, Smart Home
  - ***Test***: 2584 abstract records (2003-12) + 1095 (2013-14) SynBio (contrived from Shapira et al., 2017 search)

https://vpinstitute.org/academic-portal/tech-emergence-contest/

Key dates:

- **October-December 2018** – *CONTEST PRE-REGISTRATION*. Submit your intent to participate by completing the on-line pre-registration form with name(s) of likely participant(s) as well as your initial ideas and interests.
- **January 2019** – *CONTEST REGISTRATION and DISTRIBUTION OF PRACTICE DATASETS*. Pre-registrants will be asked to approve a data "End User Agreement" from contest data-provider Clarivate (submission deadline **February 25, 2019**). Three *practice WoS datasets* will be provided (free) to each registrant. Each practice dataset will contain WoS search results on a given domain for 12 years (e.g., ~10,000 abstract records on synthetic biology to give you 10 years of data to analyze so as to best predict activity in the following 2 years).
- **April 2019** – *CONTEST*. We will provide the *contest dataset* on **April 22nd** requiring you to send back your list of emerging topics in 10 days on **May 1st** with a brief description of your algorithm/process.
- **June 28, 2019** – *CONTEST WINNERS ANNOUNCED*
- **February – April 2019** – *CONFERENCE SUBMISSION PERIOD*. GTM2019 will be accepting abstract submissions for the *Measuring Tech Emergence* track.
- **October 17, 2019** – *CONTEST AWARD*. The winner and second prize will be awarded at the 9th Global Tech Mining (GTM2019) Conference in Atlanta, GA. *MEASURING TECH EMERGENCE SPECIAL SESSION*. We hope you all will be able to attend to as an opportunity to collectively assess and benefit from contest participant and other's research into this vital area.

Measuring Tech Emergence OVERVIEW

Measuring Tech Emergence CONTEST

Measuring Tech Emergence CONFERENCE track at GTM2019

(Evolving) Tech Emergence BIBLIOGRAPHY

Session Agenda

- Empirical Scoring – *Alan Porter and Nils Newman*
- Judges Perspectives – *Nils Newman, Dewey Murdick, Phil Shapira*
- Contest submissions:
  - Shou et al (2nd)
  - Jin Moau et al (1st)
- Panel Discussion
- Open Discussion

A notable precursor

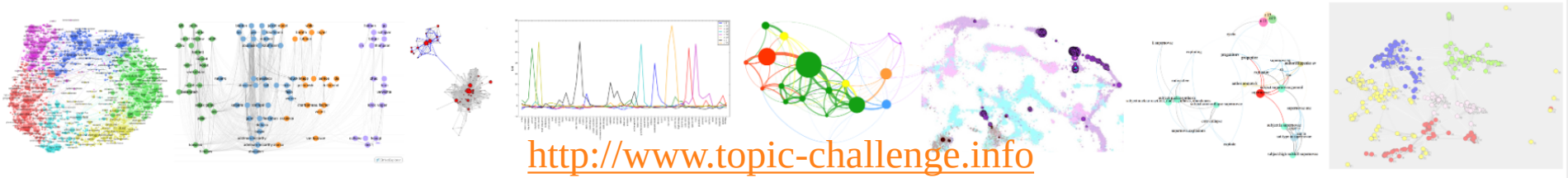
Residential Home Security Check | (68) Doug Porter - Doug's rugby | www.topic-challenge.info

Not secure | www.topic-challenge.info

Apps | Search Webmail | Google Drive | Weather | SwimOutlet.com -... | Wells Fargo | NSF FastLane | ISyE Webmail - Login | GT Library | ramblin wreck | VP Institute | Home Exercise | Shared with me - G... | Georgia Tech Library | Resurgens - Claudia

TOPIC EXTRACTION CHALLENGE

Home Participate The Data Solutions Publications About



http://www.topic-challenge.info

We invite you to join the topic extraction challenge and learn about the state of art in topic extraction in bibliometrics through systematic comparison of topic extraction approaches applied by the various groups in the field and beyond. Over the last two years, six research teams worked together to compare their approaches to the identification of thematic structures in the Astronomy and Astrophysics literature, based on a shared set of bibliographic data of 111,616 journal articles. The outcomes of this comparative exercise are published in a *forthcoming* special issue of Scientometrics. Now that Clarivate Analytics kindly agreed to make this data set available to interested researchers in the bibliometrics community, we suggest to extend this comparative approach.

We invite you to participate in the comparative topic extraction challenge!

The challenge is not to develop the best partitioning of the data set. We believe this to be impossible because there is no single best solution for two reasons. First, the structure of a body of knowledge is in the eye of the beholder, i.e. more than one valid thematic structure can be constructed depending on the perspective applied to the knowledge. Second, topical structures are reconstructed for specific purposes, so if at all, there might be a best method for a given purpose. Therefore, we challenge you to use this opportunity to gain as much information as possible about your own approach and the reasons why it produced a particular solution, and to find out how it differs from solutions produced by other approaches. We challenge you to comparatively discuss advantages and disadvantages of approaches to topic identification and thus to contribute to a cumulative body of knowledge on the suitability of data models and algorithms for the identification of topics.

How to obtain the data set is described here. Submitted solutions will be published here on this website (topic-challenge.info) and can be downloaded for comparisons. We will seek to make further tools available for comparison in the near future. If there are enough participants, we plan to run sessions on the comparative exercise at the next ISSI conferences and dedicated workshops. We hope that many of you will take up the challenge and thus contribute to cumulative progress in bibliometrics.

Kevin Boyack, Wolfgang Glänzel, Jochen Gläser, Frank Havemann, Andrea Scharnhorst, Bart Thijs, Nees Jan van Eck, Theresa Velden, Ludo Waltman
(February 2017)

5:01 PM
7/12/2019

Empirical Measures

- Using Web of Science (WoS) abstract publication records in a tech domain (~Synthetic Biology), for 2003-12 -- **predict sub-topics to be heavily researched in the following 2 years**
  - We use a threshold of terms' annualized rate of occurrence in the prediction period (2013-14)/last 2 years of the historical period (2011-12) > dataset rate of increase (27% growth), as threshold for "hits"
- Submissions to be 10 +/- 3 **terms**, or up to 10 **topics** (comprising <=10 terms each), for which we would search for occurrences in abstract records in those 2 years
- No straightforward "gold standard"

Measurement Issues

- Which fields to use?
  Title NLP (Natural Language Processing), Abstract NLP phrases, Keywords (Author), Keywords Plus [~match the submission or somewhat more inclusive level]
- "Combo Terms" combine those 4 fields
- To use exact or fuzzy match for terms? [examined both ways]
- How to handle Topics with multiple terms?

Empirical Measures

- Tallies of term frequencies in 4 periods
  - 2003-2010
  - 2011-2012
  - 2013-2014
  - 2015
- For various term fields, tailoring to each submission
- Hi-liting submitted terms/topics with **growth >= 0.27 from 2011-12 to 2013-14**
- Separating terms with lower frequencies for 2011-2014 from those with higher frequencies

High Growth Terms

- Partial illustration – alpha sorted, first 10 terms showing >0.27 growth rate, with frequency > 50 for 2011-14

	VantagePoint fuzzy matching done ~stemming				
	Combo Terms NLP >1		3	2	
	# Records:	0.27	863	1095	
# Records	Hi-lited for >=0.27, N>=50 for 2011-14	j Growth R	2011-12	2013-14	2011-14
379	application	0.32	82	108	190
320	approach	0.52	59	90	149
114	artificial cell	0.60	20	32	52
160	assembly	0.82	28	51	79
285	bacteria	0.89	55	104	159
81	biofuels	0.57	21	33	54
222	biology	0.48	50	74	124
162	biosynthesis	1.15	26	56	82
117	Biotechnology	1.44	18	44	62
689	cells	0.43	131	187	318

- High growth term hits for each submission (top row)
- Combo terms (same as previous page) – truncated list
- 1 = direct match; ~1 = approximate match
- E.g., "**metabolic engineering**" 5 hits (4 for the 13 + us)
- 17T, 20W & 16R tally ~7 or 8

5	6	topics	topics	topics	3	~7	8	4	2	8	1	0		# Records:	0.27	863	1095
2B	3C	7G	8H	11L	12M	16R	17T	18U	19V	20W	21X	22Y	# Records	Hi-lited for >=0.27, N>=50 for 2011-14	Growth R	2011-12	2013-14
													379	application	0.32	82	108
													320	approach	0.52	59	90
1													114	artificial cell	0.60	20	32
													160	assembly	0.82	28	51
													285	bacteria	0.89	55	104
													81	biofuels	0.57	21	33
													222	biology	0.48	50	74
													162	biosynthesis	1.15	26	56
													117	Biotechnology	1.44	18	44
													689	cells	0.43	131	187
													112	challenges	0.36	25	34
													193	circuits	0.52	46	70
													93	concept	0.56	23	36
													131	construction	0.34	32	43
													200	dynamics	0.63	38	62
						~1							204	E. coli	0.50	38	57
													391	effect	0.98	55	109
													159	efficiency	0.83	24	44
													276	enzymes	0.31	58	76
						1							1117	ESCHERICHIA-COLI	0.35	227	307
						1							501	gene-expression	0.41	100	141
													160	growth	0.77	27	48
													263	identification	0.74	39	68
													126	integration	0.71	21	36
													113	interest	0.44	27	39
						1							157	in-vitro	0.95	24	47
						1							143	in-vivo	1.39	15	36
													115	life	0.62	21	34
													132	living cells	0.68	25	42
													188	MAMMALIAN-CELLS	0.46	37	54
													297	mechanism	0.42	55	78
1	1						1			1			168	metabolic engineering	1.06	34	70

Similarities among Submitted Emergent Terms

- Example: 21X vs. 19V ~4 in common, shown in **bold**
- **"in common"** tally

		sum	in common		22Y	21X	20W	19V	18U	17T	16R	12M	11L	8H	7G	3C	2B
22Y	**21X**	0	22Y		X	0	0	0	0	0	0	0	0	-	0	0	0
genetic information	**genetic circuit**	24	21X			X	3	4	0	1	1	1	2	2	3	1	1
genetic information con	**Synthetic Gene Network**	26	20W				X	3	2	1	1	0	2	3	2	4	2
CaMV	**gene regulatory network or Synth**	29	19V					X	0	1	1	0	5	3	3	1	5
Petri net	**Gene Regulatory Network**	8	18U						X	1	0	0	1	1	1	1	0
B subunit	Computational Design	14	17T							X	1	0	2	1	1	1	2
adenoviral vector	Quorum Sensing	16	16R								X	0	3	2	2	0.5	2.5
actinomycete	Giant Unilamellar Vesicles/GUVs	3	12M									X	1	1	0	0	0
plant expression vector	folding DNA	20	11L										X	3	4	2	4
self-reproduction	Deletion Mutant	29	8H											X	5	3	3
IRMA	Map Kinase	28	7G												X	2	2
nonpolar residue	Supramolecular Chemistry	19.5	3C													X	2
	Metabolic Engineering	27.5	2B														X
			notes on Matrix:														
19V	18U		#s are my sense of degree of commonality --"x out of 10" aligned pretty well														
genetic circuit	heterologous expression		**Tallies are generous; looking for some commonalities**														
gene regulatory network	green fluorescent protein		Should give a quick read on how aligned or distinct a set is														
synthetic biology	cancer con																
synthetic gene	arabidopsis		**Divided by almost 2 for 16R since ~20 terms on 2 factors**														
artificial cell	directed evolution		**11L a rough look at the bigrams - very rough & divide by 2 or 3**														
systems biology	polymerase chain reaction (118 are PCR		**8H - crude estimation based on factor term emphases - so broad in 10 factors each covering 1200-4000 o**														
gene network	PCR con		**7G - also very impressionistic, as these are 6 factors with ~10 terms each (pretty broad unigrams)**														
gene therapy	molecular recognition																
genetic interaction	nucleosides con																
reverse engineering	protein kinase																

17T

- ~8 hits
- But Inclusive vs. Conservative counts can vary
- E.g., “metabolic engineering”

	Combo Terms NLP (Cleaned) (copy for 17T) - Grouped			
	Inclusive # Records	0.27	863	1095
# Records		Adj Growth Rate	2011-12	2013-14
23	cell-free protein	1.71	4	11
20	tumor necrosis	1.61	3	8
193	metabolic engineering	1.23	34	76
86	synthetic biology approach	1.11	17	36
22	heterologous gene	0.73	4	7
86	operon	0.58	17	27
44	mycoplasma	0.50	10	15
65	streptomyce	0.41	12	17
37	genetic oscillator	0.11	9	10
34	tetracycline	0.00	7	7
94	aptamer	-0.09	23	21
20	mevalonate	0.00	4	4
17	restriction site	0.00	3	3
	CONservative # Records			
168	metabolic engineering con	1.06	34	70
74	synthetic biology approach con	0.53	17	26
30	operon con	0.14	7	8
72	aptamer con	-0.06	17	16
17	genetic oscillator con	2.38	2	7
5	heterologous gene con	0.00	1	1

	Combo Terms NLP (Cleaned) - Grouped			
	# Records	0.27	863	1095
# Records		Adj Growth Ra	2011-12	2013-14
102	modules	2.31	**12**	**40**
117	Biotechnology	1.44	18	44
143	in-vivo	1.39	15	36
169	regulation	1.23	26	58
142	plants	1.16	18	39
162	biosynthesis	1.15	26	56
149	strains	1.14	27	58
168	metabolic engineering	1.06	34	70
95	prediction	1.05	19	39
391	effect	0.98	55	109
157	in-vitro	0.95	24	47
285	bacteria	0.89	55	104
159	efficiency	0.83	24	44
149	understanding	0.83	24	44
160	assembly	0.82	28	51
401	production	0.80	67	121
160	growth	0.77	27	48
263	transcription	0.76	42	74
263	identification	0.74	39	68
126	integration	0.71	**21**	**36**
132	living cells	0.68	25	42
200	dynamics	0.63	38	62
115	life	0.62	21	34

Two of the Topic Submissions

7G TEXT FIELDS: Abstracts

Rank	Topic	terms
1	Topic 1	biological,use,gene,system,design,cell,engine,model,molecular,synthetic
2	Topic 6	DNA, base, bind, structure, acid, pair, protein, differ, RNA, oligonucleotide
3	Topic 2	network, circuit, synthetic, genetic, dynamic, model, system, biological, method, design
4	Topic 5	DNA, use, detect, sequence, oligonucleotide, target, method, assay, synthetic , probe
5**7G**	Topic 9	cell, membrane, surface, use, lipid, artificial, metabolism, pathway, particle, vaccine
6	Topic 4	gene, express, protein, transcript, cell, synthetic, function, pathway, active, interact

IELDS: Title and Abstract

Submission – Table 2

topic	term
1	biolog
	synthet
	system
	design
	model
	genet
	cell
	engin
	circuit
	base
2	network
	gene
	model
	synthet
	genet
	design
	biolog
	robust
	oscil
	method

Another Topic Submission

S: Title and Abstracts

Topic	Topic words (s	emerging(2-gram) EMERGENT TERMS
1	biology	"synthetic biology", "biology application", "system biology", "molecular biology"
	synthetic	"synthetic biology", "synthetic gene", "synthetic biological", "application synthetic", "synthetic promoter", "field synthetic", "synthetic biologist", "engineere synthetic", "system synthetic", "construction synthetic", "synthetic dna", "synthetic oligodeoxynucleotide", "synthetic system", "synthetic sequence", "gene synthetic", "synthetic genome", "express synthetic", "optimize synthetic", "synthetic circuit", "expression synthetic", "development synthetic", "synthetic cell", "synthetic oligonucleotide", "construct synthetic", "short synthetic", "synthetic fragment", "synthetic network", "synthetic oligonucleotides", "synthetic molecule", "model synthetic", "synthetic genetic"
	engineere	"metabolic engineere", "engineere biological", "engineere synthetic", "reverse engineere", "genetic engineere", "tissue engineere"
	system	"biological system", "genetic system", "expression system", "model system", "system synthetic", "synthetic system", "immune system", "delivery system", "control system", "molecular system", "system biology", "component system", "cell system"
	biological	"biological system", "synthetic biological", "biological network", "engineere biological", "biological function", "biological active", "biological process"
	genome	"synthetic genome", "genome wide"
	assemle	
	recent	
	field	"field synthetic"
	development	"development synthetic"
2	enzyme	"restriction enzyme"
	production	"protein production"
	strain	"coli strain"
	substrate	"substrate specificity"
	produce	
	alpha	
	pathway	"metabolic pathway", "biosynthetic pathway", "signale pathway"
	gene	"gene expression", "synthetic gene", "gene cluster", "gene circuit", "gene network", "gene therapy", "essential gene", "yeast gene", "gene deletion", "gene regulatory", "encode gene", "gene silenc", "gene construct", "target gene", "gene promoter", "gene encode", "identify gene", "gene synthetic", "expression gene", "gene clone", "gene synthesis", "control gene", "gene require", "reporter gene", "gene express", "gene identify", "level gene", "gene product", "gene function", "thymulin gene", "gene involv", "express gene", "gene carrier", "gene sequence", "gene code", "gene delivery", "gene transfer", "gene regulation", "multiple gene"
	activity	"promoter activity", "enzymatic activity"
		"escherichia coli", "coli codon", "coli strain", "coli cell"

Session Agenda

- Empirical Scoring – *Alan Porter and Nils Newman*
- Judges Perspectives – *Nils Newman, Dewey Murdick, Phil Shapira*
- Our 2 contest standouts:
  - Prof. Zhengyin Hu – Shou et al–Beijing University of Technology (2nd)
  - Jin Mao et al –Wuhan University (1st)
- Panel Discussion
- Open Discussion

Team	text fields	data supplementation	#terms	type	Algorithm
AIT Austrian Institute of Technology GmbH Edgar Schiebel	Titles , Abstracts, Keywords-Author, Keywords-Plus	no	13 terms	noun-phrases; includes acronyms	time series, MS Access tables with calculations for novelty, growth, applicability, interdisciplinarity
Wuhan University Chao Mao	Titles, Abstracts, Keywords-Authors	PubMed Mesh	10 terms	noun-phrase	temporal exponential random graph model (ERGM); bibtex
Wuhan University Jin Mao,	Titles, Abstracts, Keywords-Author, Keywords-Plus	WoS references (assume reference, citations and fund sponsors as they are used in calculation)	13 terms	noun-phrases	neural network based solution; The Termolator (open source tool)
Beijing University of Technology Shuo Xu	Titles and Abstracts	WoS cited references according to DOIs	10 terms	noun-phrases	TNG (topical n-grams) model
Fudan University Li Tang	Titles, Keywords-Authors, Keywords-Plus	reprint author affiliation information from WoS	10 terms	noun-phrases	VantagePoint
Nanjing University Chao Min, (Tao Han)	Titles and Abstracts	no	10 topics	10 stemmed words (single) per topic; includes acronyms	Delay index and boost value (For every single term, we construct its time series. Taking advantage of these time series data, we select those most emerging terms based on one of our models. And then LDA is applied to those terms to find semantic topics. At last we select 10 (or less) most emerging topics on the basis of the terms these topics include)
Chengdu library and Information Center, Chinese academy of Science Yan Qi	Titles and Abstracts	no	10 topics	varying-sized list of "emerging(2-gram)" s per topic (terms in Topic Word column to be ignored)	Python's Gensim toolkit; LDA topic recognition with optimized parameters