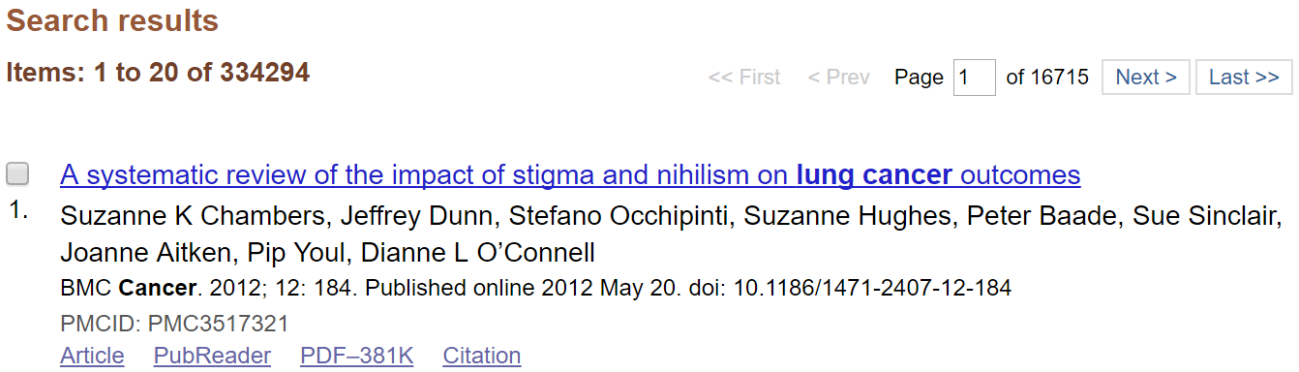
Building Content-driven Entity Networks for Scarce Scientific Literature using Content Information

Reinald Kim Amplayo

Min Song

Biomedicine and big data

- Big data is now everywhere (and is available and cheap)
  - In the biomedical field – research papers, clinical data, etc.
- Big data is used in almost anything
  - Biocuration
  - Entity extraction
  - and many more knowledge discovery tasks
- For example: lung cancer papers from PMC



Search results

Items: 1 to 20 of 334294

<< First < Prev Page 1 of 16715 Next > Last >>

1. A systematic review of the impact of stigma and nihilism on **lung cancer** outcomes

Suzanne K Chambers, Jeffrey Dunn, Stefano Occhipinti, Suzanne Hughes, Peter Baade, Sue Sinclair, Joanne Aitken, Pip Youl, Dianne L O'Connell

BMC **Cancer**. 2012; 12: 184. Published online 2012 May 20. doi: 10.1186/1471-2407-12-184

PMCID: PMC3517321

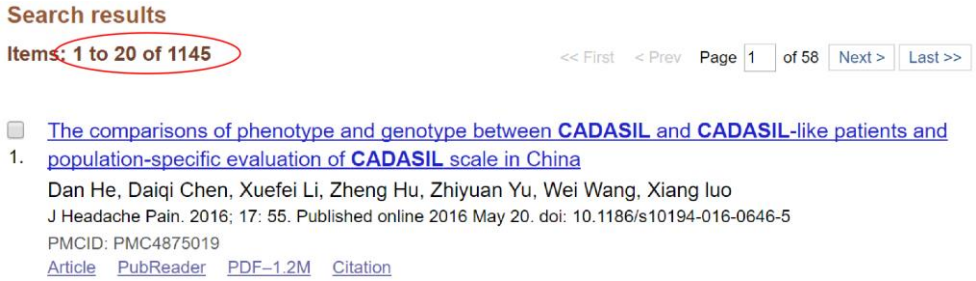
Article PubReader PDF–381K Citation

Knowledge discovery using big data

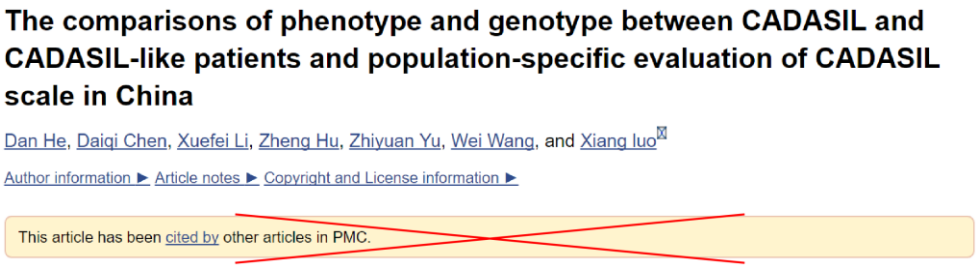
- One way to discover knowledge from big data is through network construction
  - Finding **prolific authors** using author collaboration (Hou et al., 2008) networks
  - Determining **strength of authors** using author co-citation (Ding, 2011) and author citation (Zyczkowski, 2010) networks
  - Finding important **biological entities** and **keywords** (Plake et al., 2006; Ding et al., 2013), and **topics** (Lee et al., 2016) using entity co-occurrence and entity citation networks
  - Finding **author communities** (Song et al., 2014), **topical communities**, etc.
- Two major kinds of social/knowledge networks
  - Entity = {author, bio-entity, keyword, topic, …}
  1. **Entity co-occurrence networks** – given a scope (abstract, author list, etc.) with two or more entities, connect all possible pairs
  2. **Entity citation networks** – given scope A with links to other scope (document citing another document), connect

How about using small (scarce) data?

- Scarcity?
  - Lack in **volume**: data size is not big enough to discover knowledge



  - Lack in **value**: information (author list, abstract, …) needed is not found



- Examples
  - Rare diseases: CADASIL, arthrogryposis, etc.
  - Genes: NOTCH3, RBPJ, etc.

Methodology (0/1)

- Two-in-one solution
  - Instead of using the meta information (author list, abstract, citation information), **Use full-text** content information to extract entities and construct the networks!
- Collaboration/co-occurrence networks
  - Traditional way: use the given author list (for author collaboration) and abstract (for co-occurrence) meta information
  - Full-text content: use the authors in the **reference section** (for author collaboration) and the **in-text citation sentences** (for co-occurrence)
- Citation networks
  - Traditional way: use the citation meta information, combined with the author list or abstract meta information
  - Full-text content: use the paper's **authors** and **abstract** to extract citing entities and use the authors in the **reference section** and the **in-text citation sentences** to extract cited entities

Methodology (1/1)

- Traditional methods only use A; our methods utilize A and B.

Metadata

The comparisons of phenotype and genotype between CADASIL and CADASIL-like patients and population-specific evaluation of CADASIL scale in China

Dan He, Daiqi Chen, Xuefei Li, Zheng Hu, Zhiyuan Yu, Wei Wang, Xiang luo

Background: Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is the most common form of hereditary stroke disorder caused by mutations in the NOTCH3 gene. Although CADASIL scale is a widely used tool to screen clinically suspected CADASIL patients, the differential effects of this scale in various populations remain unknown. Methods: 92 CADASIL-like patients and 24 CADASIL patients were selected based on CADASIL scale and gene tests. The clinical, genetic and radiological characteristics were analyzed. Results: Based on the CADASIL scale, we first screened 116 suspected CADASIL patients, and detected 20 mutations in 24 CADASIL-patients (Specificity: 20.69 %). Surprisingly, we found that transient ischemic attack/stroke, migraine, cognitive decline, psychiatric disturbances and early onset age in CADASIL scale showed no differences between the CADASIL and the CADASIL-like patients ($p>0.05$). Instead, recurrent cerebral ischemic events (58.33 %, $p=0.028$) and positive family histories ($p<0.05$) were more frequently observed in CADASIL patients. Moreover, compared with CADASIL-like patients (21.74 %), CADASIL patients demonstrated higher percentage of temporal pole involvements (58.33 %, $p=0.001$), but not the external capsule involvements (66.67 %, $p=0.602$), in MRI imaging. Further, we found that vascular risk factors could occur in both CADASIL patients and CADASIL-like patients, and therefore could not be used as the markers to differentiate the two groups in our study ($p>0.05$). By performing DSA analysis, we for the first time identified dysplasia of cerebral blood vessels in CADASIL patients, which were detected more frequently in CADASIL patients (41.67 %) in comparison with CADASIL-like patients (8.69 %, $p<0.01$). Conclusion: Our data suggested that the efficacy of CADASIL scale to diagnose the disease varied with specific populations. Recurrent cerebral ischemic events, temporal pole involvements (but not the external capsule) in MRI imaging and dysplasia of cerebral blood vessels in DSA may be the new potential risk factors of the CADASIL scale suitable for Chinese patients. Gene testing by encephalopathy gene panel is expected to improve the accuracy of CADASIL differential diagnosis and increase the understanding of this disease in the future.

A

Full Text

Introduction Go to:

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is the most common form of hereditary small vessel disease (SVD), and is linked to mutations in the *NOTCH3* gene [1]. The clinical features in CADASIL are characterized by recurrent strokes, migraine with aura, motor deficits, pseudobulbar palsy, mood disturbances and subcortical dementia [1]. The profile of cognitive impairment in CADASIL resembles that in sporadic vascular cognitive impairment (VCI), and manifests as deficits in attention, processing speed and executive function, but relatively preserved semantic fluency [2]. CADASIL subjects exhibit rather specific spatial distribution of white matter (WM) changes as shown by magnetic resonance imaging (MRI) suggesting disrupted cortical connectivity underlies the cognitive deficits. Abnormalities in normal-appearing WM are not readily demonstrable with conventional MRI, but become visible with diffusion tensor imaging (DTI) or magnetization transfer imaging. However, WM hyperintensities on normal MRI did not correlate with cognitive dysfunction in CADASIL [3]. In contrast, DTI was shown to relate to impairment in executive function in SVD as well as CADASIL [4,5]. Furthermore, DTI histogram metrics were used to predict disease progression in CADASIL [6,7].

⋮

References Go to:

1. Chabriat H, Joutel A, Dichgans M, Tournier-Lasserve E, Bousser M-G. CADASIL. Lancet Neurol. 2009;8:643–653. [PubMed]

2. Buffon F, Porcher R, Hernandez K, Kurtz A, Pointeau S, Vahedi K, Bousser MG, Chabriat H. Cognitive profile in CADASIL. J Neurol Neurosurg Psychiatry. 2006;77:175–180. [PMC free article] [PubMed]

B

Advantages

- **Larger communities**
  - Data is now relatively bigger than before
- **Clearer polarity**
  - Reflection of citations on co-occurrence networks -> redundant edge weighting
  - Results to much clearer distinction between important and unimportant nodes
- **The use of citation information**
  - Entities in abstracts = entities cited + entities not cited
  - Extracts only entities that are cited (which may not be found in abstracts)
- Disadvantage?
  - Data is dirty due to automated entity extraction
  - However, big data covers the issue (more data -> better accuracy)

Experimental setting (0/1)

- Use case: CADASIL (scarce) and Metformin (pseudo-scarce)
- Dataset
  - CADASIL papers found in PubMed Central (using the query *cadasil*)
  - 1000 metformin papers found in PubMed Central (using the query *metformin*)

		author	bio	keyword	topic
traditional co-occurrence	nodes	4,707	3,493	17,033	-
	edges	18,948	40,386	369,818	-
our method co-occurrence	nodes	84,180	21,897	142,319	-
	edges	295,066	89,298	846,269	-
our method citation	nodes	87,719	24,522	150,895	498
	edges	952,994	310,590	4,513,469	17,603

Experimental setting (1/1)

- Networks constructed
  - Three kinds: traditional co-occurrence, our-method co-occurrence, our-method citation
  - Traditional citation network cannot be constructed due to scarcity (lack of value)
  - Four entities: authors, biological entities, keywords, topics
- Node ranking
  - PageRank (Page et al., 1999) with $\delta = 0.5$, following (Chen et al., 2007)
- Entity extraction methods
  - Authors: ABNER (Settles, 2005)
  - Biological entities: PKDE4J (Song et al., 2015)
  - Keywords: RAKE (Rose et al., 2010)
  - Topic: LDA (Blei et al., 2003)

Finding prolific authors

- Compared PageRank ranking to (1) author's h-index and (2) the quotient of total citations over the number of documents of the author, c/d

(a) traditional co-occurrence

Author	h	c/d
HS MARK...	76	62.45
TR BARR...	29	38.09
AJ LAWR...	39	21.49
RG MORR...	61	46.19
M TRAYL...	10	19.31
C LAMBE...	8	14.38
P BENJA...	2	1.88
RL BROO...	7	9.64
S BEVAN...	22	40.41
B PATEL	8	9.45
average	*26.2*	*26.33*

(b) full text-based co-occurrence

Author	h	c/d
A JOUTEL	41	92.47
E TOURN...	57	59.28
MG BOUS...	87	58.19
H CHABR...	56	46.56
K VAHEDI	36	73.19
V DOMEN...	16	162.88
MM RUCH...	26	39.86
J WEISS...	112	154.07
E MAREC...	25	22.61
EA CABA...	23	13.64
average	47.9	**72.27**

(c) full text-based citation

Author	h	c/d
H CHABR...	56	46.56
A JOUTEL	41	92.47
MG BOUS...	87	58.19
M DICHG...	58	40.64
E TOURN...	57	59.28
K VAHEDI	36	73.19
HS MARK...	76	62.45
N PETERS	24	34.43
F FAZEK...	77	44.16
JM WARD...	71	34.00
average	**58.3**	54.54

Finding important bio-entities

- For simplicity, extracted only genes and diseases
- Compared top ranked bio-entities to MalaCards (Rappaport et al., 2013)

traditional co-occurrence	**notch3**, **vascular dementia**, **stroke**, **hypertension**, alzheimer's disease, **migraine**, disease, vascular lesion, **ischemia**, **notch1**, multiple sclerosis, **amyloid angiopathy**, lacunar infarct, diabetes, single gene disorder, genetic disorder, **atherosclerosis**, allele, vascular, cortex
full text-based co-occurrence	**notch3**, **notch1**, **notch2**, **stroke**, alzheimer's disease, **hypertension**, multiple sclerosis, **vascular dementia**, dll4, **jag1**, **ischemic stroke**, **amyloid angiopathy**, **migraine**, disease, dll1, **fabry disease**, human disease, carasil, lacunar stroke, **atherosclerosis**
full text-based citation	**notch3**, **stroke**, **hypertension**, **caa**, alzheimer's disease, **notch1**, **migraine**, **atherosclerosis**, **vascular dementia**, lacunar infarct, disease, vascular lesion, **cvd**, diabetes, **notch2**, cortex, **ischemia**, dll4, skin, **brain atrophy**

Finding meaningful keywords

- RAKE automatically extracts keywords from the text – some keywords may not be related to the topic
- Compared top ranked keywords to MalaCards

traditional co-occurrence	homonymous visual field defect, **small vessel disease**, **vascular disease**, central retinal artery occlusion, intracranial pressure, optic disc edema, ischemic optic neuropathy, homonymous hemianopia, external carotid artery, ocular ischemic syndrome, **visual loss**, spontaneously, retinal ischemia, optic tract, retinal infarction, **cerebral white matter**, **central nervous system**, clinical presentation, cerebral atrophy, blood flow
full text-based co-occurrence	**cadasil**, **subcortical infarct**, **notch signaling**, risk factor, **vascular dementia**, **cognitive impairment**, **notch receptor**, **cerebral amyloid angiopathy**, **multiple sclerosis**, alagille syndrome, endothelial cell, **stroke**, **notch pathway**, **notch**, **alzheimer disease**, **cognitive decline**, risk, **notch signaling pathway**, disease, **small vessel disease**
full text-based citation	**notch signaling**, **cognitive impairment**, risk factor, endothelial cell, **cognitive decline**, **white matter**, risk, **alzheimer disease**, **notch receptor**, cognitive function, **cadasil**, cell, **stroke**, **subcortical infarct**, **ischemic stroke**, evidence, **notch**, **vascular risk factor**, previously, **notch signaling pathway**

Discovering influential topics

- Assumed one snippet of text (abstract, in-text citation sentence) has only one major topic -> experiment can only be done in entity citation networks
- Top 5 influential topics about CADASIL

Topic 443	Topic 297	Topic 461	Topic 243	Topic 361
risk	cell	study	matter	study
factor	notch	disease	disease	matter
diabetes	stem	research	svd	brain
hypertension	signaling	approach	lesion	impairment
smoking	differentiation	datum	wmh	association
disease	progenitor	treatment	stroke	lesion
stroke	fate	review	lacunar	mri
study	development	result	hyperintensity	volume
age	pathway	patient	vessel	wmh
mellitus	role	disorder	mri	wml

Pseudo-scarce data with metformin drug

- Ding et al. (2013) constructed a traditional entity citation network using all available papers regarding the metformin drug in PMC
- We compared the above network to our method using only 1000 papers out of all the papers available

out-degree citation (Ding et al., 2013)	traditional co-occurrence	full text-based co-occurrence	full text-based citation
insulin	oglcnac	**slc2a4**	**slc2a4**
large	p78	gene	gene
impact	p180	**sirt1**	**sirt1**
lep	p202 ptp1b gene	**nfe2l2**	**nfe2l2**
tnf	trem1	met	ae
renin	**slc2a4**	glp1 ras	ppg
insulin receptor	dpp4	ppg	met
set	**pparg**	**tp53**	**pten**
mmp9	sglt2	ae	**tp53**
mmp2	ae	**pten**	sglt2

Conclusion

- Proposed an improved method to constructing social and knowledge networks using content information
  - Advantages are three-fold: larger communities, clearer polarity, and citation emphasis
- Did experiment on CADASIL data and constructed networks using four entities to (a) find prolific authors, (b) find important bio-entities, (c) find meaningful keywords, and (d) discover influential topics
  - Our method performed significantly better than the traditional methods
- Compared our method to traditional methods using big data
  - Our method is comparable to or better than the traditional methods, even with unfair amounts of data

Thank you!

- If you have questions, ask them during poster sessions
- References are found in the paper
- Thank you!

- Text and Social Media Mining Lab
- Yonsei University, Seoul, Korea