Advanced Large Scale Cross Domain Temporal Topic Modeling Algorithms to Infer the Influence of Recent Research on IPCC Assessment Reports

Jennifer Sleeman[1], Dr. Milton Halem[1], Dr. Tim Finin[1], Dr. Mark Cane[2]
(1) Dep't of Computer Science and Electrical Engineering, University of Maryland, Baltimore County (UMBC)
(2) Lamont-Doherty Earth Observatory, Columbia University

ABSTRACT

One way of understanding the evolution of science within a particular scientific discipline is by studying the temporal influences that research publications had on that discipline. We provide a methodology for conducting such an analysis by employing cross-domain topic modeling and local cluster mappings of those publications with the historical texts to understand exactly when and how they influenced the discipline. We apply our method to the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports and the citations therein. The IPCC reports were compiled by thousands of Earth scientists and the assessments were issued ~every five years over a 30 year span, and includes over 200,000 research papers cited by these scientists.

INTRODUCTION

Given a particular field of scientific study that has a large impact on society, or extreme events that affect the world on a global level, committees or panels are often formed to make reports and recommendations to governments or global enterprises based on the collective work of many individual experts and on large bodies of scientific literature and observational data. Such reports often convey findings that are based on years of related published work augmented in the voluminous numbers of citations accompanying these reports.

Can we build a machine intelligence system that integrate large multi-faceted conceptual text content reports with massive numbers of cited documents to form conclusions that are either consistent with human-formed conclusions, or perhaps better?

Can we quantify which cluster of referenced citations strongly contributed to or influenced a finding or recommendation by understanding the prior relationships between the report domain and the citation domain?

Concretely, can we build a machine intelligent system that models text in such a way that researchers or authors can discover relevant citations that could improve the preparation of the IPCC assessment reports?

METHODS

Our methodology consist of:

- performing text conversions
- citation retrieval
- text pre-processing
- model generation
- model correlation and document clustering

Each report is a a combination of books and within each book is a set of chapters with subsections. We formalize this below.

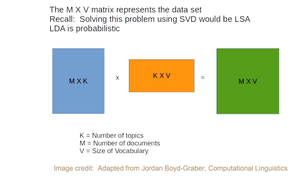
There are n reports $ar_1, ar_2, ..., ar_n$, currently $n = 5$.

There are m books $br_{n,1}, br_{n,2}, ...br_{n,m}$ where $br_{n,m} \subset ar_n$, currently $m = 4$ for all ar_n.

There are l chapters $ch_{n,m,1}, ch_{n,m,2}, ...ch_{n,m,l}$ where $ch_{n,m,l} \subset br_{n,m}$.

TECHNICAL DETAILS

Topic modeling algorithms are used to find latent variables or 'topics' that describe the thematic structure of a collection of documents. Topic modeling is based on early seminal work by Deerwester et al. who introduced the concept of Latent Semantic Analysis which uses singular value decomposition resulting in a document to term matrix.



LSA

Image credit: Adapted from Jordan Boyd-Graber, Computational Linguistics I: Topic Modeling, 2013

LDA

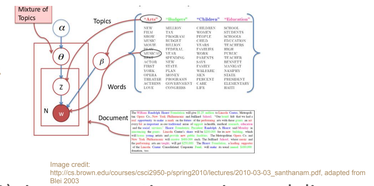


Image credit: http://cs.brown.edu/courses/csci2950-p/spring2010/lectures/2010-03-03_santhanam.pdf, adapted from Blei 2003

Latent Dirichlet Allocation (LDA) (Blei 2003) is a generative topic modeling method that generates 'topics' showing a statistical relationship between observed and latent random variables. Each topic is a probability distribution over a collection of words. Each document is a mixture of topics. Topics are drawn from a Dirichlet distribution. Inference is performed by using variational and sampling methods such as Gibbs Sampling.

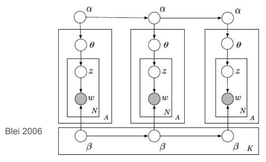
$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right)$$

Blei 2003

Joint distribution marginal probability of observed variables (computed by summing the joint distribution over all possible hidden topic structures)

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}.$$

Where: β_k topics,
θ_d topics proportions,
z_d topics assignments,
w_d observed word for doc *d*



Blei 2006

Topics and topic proportions are chained together to evolve over time

Extensions to topic modeling algorithms have included temporal components such as discrete dynamic topic modeling (dDTM) (Blei 2006) and continuous topic modeling (cDTM) (Wang et al. 2012).dDTM assumes latent topics change over the collection and assumes the timestamps of the documents in the collection can be modeled using a Gaussian random walk process.

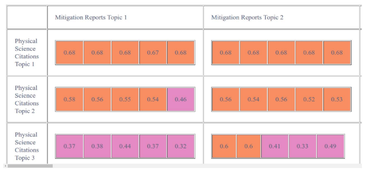
Citation Parsing & Retrieval

Book	Precision	Recall	F-Measure	Total Retrieved
Physical Science	.91	.73	.81	~53000
Mitigation	.94	.58	.72	~77000
Impact	.91	.74	.82	~18600
Synthesis	.99	.81	.89	~60

Example Dynamic Topic Model

AR1	AR2	AR3	AR4	AR5
climate change	climate change	climate change	climate change	adaptation
radiative forcing	radiative forcing	kyoto protocol	adaptation	climate change
temperature	temperature	clean development mechanism	temperature	global mean surface temperature
kyoto protocol	kyoto protocol	radiative forcing	radiative forcing	surface temperature
land use	land use	temperature	kyoto protocol	equilibrium climate sensitivity
adaptation	adaptation	land use	land use	carbon dioxide
clean development mechanism	clean development mechanism	adaptation	equilibrium climate sensitivity	temperature
greenhouse gas	greenhouse gas	global warming	surface temperature	land use
global warming	global warming	climate model	clean development mechanism	anthropogenic
climate model	climate model	greenhouse gas	climate model	radiative forcing

Example Cross-Domain Divergence

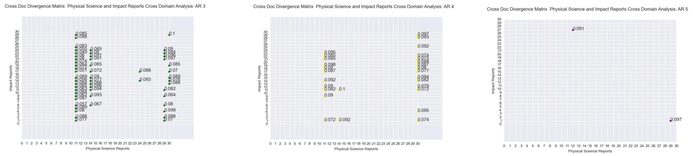


EARLY RESULTS

We experiment with three different domain influence models. In each case we build dynamic topic models for each domain, we create micro-domains and build divergence tensors. We set a threshold for the divergences of interest and pair the domain topics based on this subset of divergences. Given a document topic threshold, we find documents related to the topic of interest.

We validated our method using accuracy as a function of thresholding for finding the citations that are currently referenced in the reports. For this test we used Physical Science citations as domain 1 and Physical Science reports as domain 2.

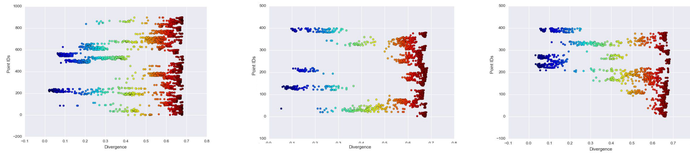
Threshold	Accuracy
.05	65%
.10	43%
.25	26%
.50	15%
.75	4%



Physical Science Citations & Impact Reports

Physical Science Citations & Mitigation Reports

Impact Citations & Mitigation Reports



Domain 1 Physical Science Citations

AR	Chapter	Title
3	13	Bringing a Global Issue Closer to Home
4	10	Emulation and probabilistic climate predictions
5	10	Emulation and probabilistic climate predictions

Domain 2 Impact Reports

AR	Chapter	Title
2	18	Human Population Health
3	9	Human Health
4	8	Human Health
5	11	Human Health

Our method was able to find a citation in the Physical Science book that has a strong correlation to human health as the paper discusses 'consequences for wildlife','changes in water supply and quality', 'political and economical impact' and other topics that are strongly related to humans and their health. This citation was not cited under any chapter in the Impact book. It should be noted that there are no chapters in the Physical Science book specific to 'Human Health'.

CONCLUSIONS

Our mapped climate change models of citation and report domains conveys how citations may influence reports. Understanding influence provides two main benefits. It can be used to make predictions for the next IPCC assessment report. We have shown it could be used to discover relevant citations that were not referenced or cited.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
D. M. Blei and J. D. Lafferty. "Dynamic Topic Models". In Proceedings of the 23rd International Conference on Machine Learning, 2006.
Wang, Chong, David Blei, and David Heckerman. "Continuous time dynamic topic models." *arXiv preprint arXiv:1206.3298* (2012).
Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane, "Machine Learning the Temporal Evolution of Climate Change: Dynamic Topic Models for Concept Clustering Using Cross-Domain Divergence Maps", Submitted to AAAI Conference 2017
Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane, "Dynamic Topic Modeling to Infer the Influence of Research Citations on IPCC Assessment Reports", Submitted to IEEE BigData 2016
Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane, "Advanced Large Scale Cross Domain Temporal Topic Modeling Algorithms to Infer the Influence of Recent Research on IPCC Assessment Reports", Submitted Abstract to 2016 AGU Fall Meeting