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

Jennifer Sleeman¹, Dr. Milton Halem², Dr. Tim Finin³, Dr. Mark Cane²

(1) Dep’t of Computer Science and Electrical Engineering, University of Maryland, Baltimore County (UMBC)
(2) Lamont-Doherty Earth Observatory, Columbia University

ABSTRACT

Our methodology consists of:

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

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 consists of:

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

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.

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.

\[ \begin{align*}
&\Phi, \Theta, Z, \omega, \nu, \xi \\
&= \max_{\Phi, \Theta, Z, \omega, \nu, \xi} \sum_{d=1}^{D} \sum_{z=1}^{K} \omega_{dz} \log(\Phi_{z}^{T} \Theta_{d}) \\
&= \sum_{d=1}^{D} \sum_{z=1}^{K} \omega_{dz} \max_{\Phi, \Theta, Z, \omega, \nu, \xi} \log(\Phi_{z}^{T} \Theta_{d})
\end{align*} \]

\[ \begin{align*}
\text{Joint distribution marginal probability of observed variables} \ &\text{computed by summing the joint distributions over all possible hidden topic structures} \\
\text{Extensions to topic modeling algorithms have included temporal components such as discrete dynamic topic modeling (dDTM) (Blei 2006) and continuous topic modeling (cDTM) (Wang et al. 2012). dDTM assumes latent topics change over time and the cDTM allows topics and their assignments to change over time.}
\end{align*} \]

EARLY RESULTS

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.

CONCLUSIONS

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 document topics to the report topics. 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 task we used Physical Science citations as domains 1 and Physical Science reports as domain 2.