Publication Mining using Social Network Analysis and Probabilistic Topic Modeling

Supriya Chinthavali[1], Dakotah Maguire[2], Ethan Hicks[3], Philip Hicks[3]
Oak Ridge National Laboratory[1], ORAU[2], HERE[3]

ABSTRACT

The amount of scientific information available on-line has increased at an exceptional rate, with recent estimates reporting that a new paper is getting published every 20 seconds.

The 4 main challenges in the area of publication mining are
1.Identify the individual or group of individuals who appear to be the expert in a particular field or subfield.
2.Identify topics that have been researched across all publications
3. Visualize the geographic distribution of the topics in the publications
4. Identify how topics have shifted over time

The data mimers team (ORNL+UTK) provides solutions to the four challenges mentioned above using **machine learning** and graph modelling techniques called and social network analysis **probabilistic topic models** to mine the scientific publications and identify key characteristics and patterns that can be used by human researchers to develop useful knowledge and further enhance scientific discovery.

Our approach involves:
1.Constructing various **collaboration/citation networks and** enable analysis of graph-based collaboration networks that can identify experts within a given area.
using computing community detection algorithms to answer analytic questions posed by this data challenge.
2.Developing a **visual–analytic** tool that enables a user to understand the key topics that have been researched over the previous 50 years using an unsupervised bayesian probabilistic model known as Dynamic Topic Model based on LDA(Linear Dirichlet Allocation).

POINTS OF CONTACTS

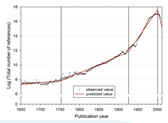
Supriya Chinthavali

Data Visualization and Machine Learning Team,
Computational Data Analytics Group,
Computational Sciences and Mathematics Division,
Oak Ridge National Laboratory
(865) 574-7592
chinthavalis@ornl.gov

INTRODUCTION

SCIENTIFIC PUBLICATION MINING USING TOPIC MODELLING AND SOCIAL NETWORK ANALYSIS

ORGANIZE SEARCH
VISUALIZE PREDICT
SUMMARIZE UNDERSTAND



According to bibliometric analysts Lutz Bornmann, at the Max Planck Society in Munich, Germany and Ruediger Mutz, global scientific output **doubles** roughly every nine years **.**

Topic modeling is an emerging field in machine learning, which allows us to analyze streaming collections such as publications, see how the discovered themes are connected to each other, and how they change over time. Topic models are also referred to as probabilistic topic models, which refers to statistic algorithms for discovering the latent semantic structures of an extensive text body. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies.

Social Network Analysis is based on social networks that are derived from interactions between individuals(collaborations, relationships etc) and thereby model meaningful social relationships. There is a natural inclination in a social group to get associated to respected members of the groups, and hence respected members of a community will tend to be connected to the majority of the community members either directly or indirectly via others. Thus, social networks could provide valuable insights on key experts within publications related network.

CHALLENGES AND OUR APPROACH

CHALLENGES


Forbes
http://onforb.es/zzIZw9y
DataScience Platform - For Data Scientists
A data science platform that empowers data scientists to just do data science datascience.com/platform
5/23/2016 @ 8:30AM | 31,120 views
Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Processing time and memory issues

PROBABILISITIC TOPIC MODELLING AND SOCIAL NETWORK ANALYSIS

Social Network Analysis
Collaboration Networks (direct citation, co-citation, co-author networks)

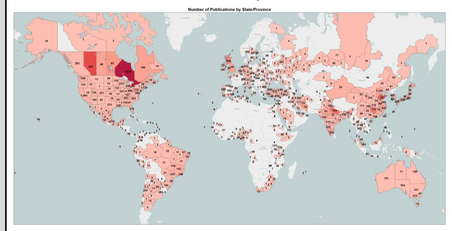
Dynamic Topic Modelling (LDA based model)

Network Visualization using Cystoscope and Information vis using Tableau

Our approach involves:
1.Construction of several **collaboration based networks** using citation and authors data and applying community detection algorithms to answer analytic questions posed by this data challenge.
2.Developing a **visual–analytic** tool that enables a user to organize the documents that have been researched over the previous 50 years based on the underlying theme structure and also provide graph-based collaboration networks that can identify experts within a given area.

VISUALISATION:GEOGRAPHIC DISTRIBUTION

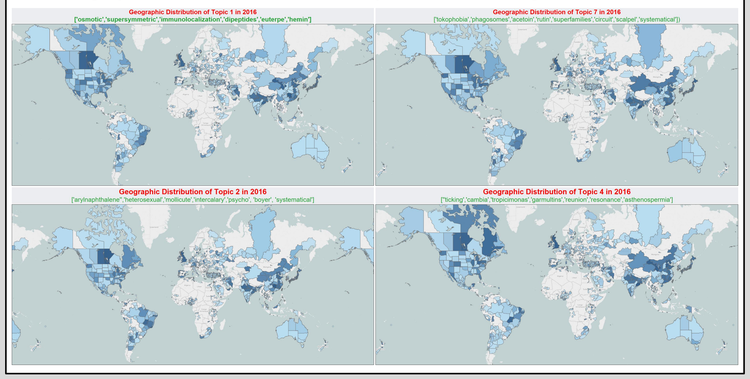
Visualization of number of publications for 2016



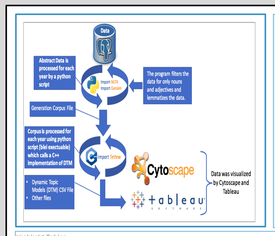
Using R, we take the output files of DTM (gam.dat) and combine the probabilities of words in each topic to get topic distribution of the documents. We then join this with the affiliation data to plot the geographic distribution of topics on a world map.

Note: Due to missing or uninterpretable data (e.g. various admin levels), 7.5% of the publications for 2016 were not geocoded.

Geographic Distribution of various topics detected using DTM for 2016

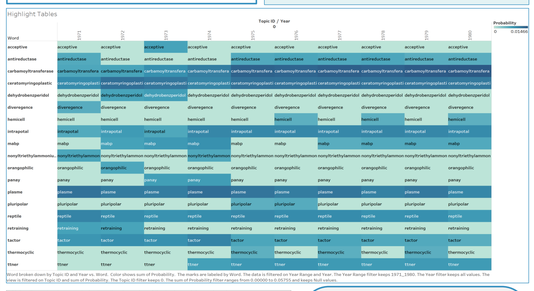


TOPIC MODELLING



The **Latent Dirichlet allocation** (LDA) is the simplest statistical topic model. The intuition behind LDA is that documents exhibit multiple topics. A topic is defined to be a distribution over a fixed vocabulary. Each document exhibits the topics in different proportion; each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics.

A **dynamic topic model (DTM)** considers the ordering of the documents and provides a better posterior topical structure than LDA[2].Rather than a single distribution over words, a topic is now represented as a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time.



Using the natural parameterization, the dynamics of the topic model are given by

$$\beta_{k,t}|\beta_{k-1,t} \sim N(\beta_{t-1,k}, \sigma^2 I)$$

and

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I).$$

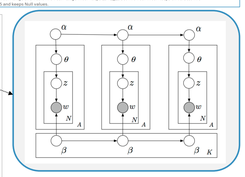
Graphical Model for DTM

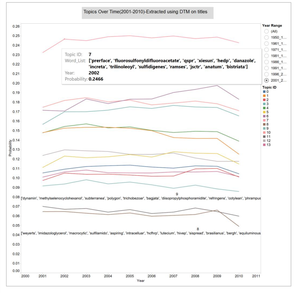
The generative process at time slice 't' is therefore:

1. Draw topics $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \forall k$
2. Draw mixture model $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$
3. For each document:
   1. Draw $\eta_{t,d} \sim N(\alpha_t, a^2 I)$
   2. For each word:
      1. Draw topic $Z_{t,d,n} \sim \text{Mult}(\pi(\eta_{t,d}))$
      2. Draw word $W_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z_{t,d,n}}))$

where $\pi(\sigma)$ is a mapping from the natural parameterization x to the mean parameterization, namely

$$\pi(x_t) = \frac{\exp(x_t)}{\sum_i \exp(x_t)}.$$





After executing DTM (C++ impementation by Dr David Blei[2]), a list of files get generated:

1. **topicxxxvarelogprob.dat**: this file presents the distribution of words for the topic xxx for each period analyzed. This values are in logarithm.
2. **gam.dat** : this file stores the parameters of the Dirichlet variational (gammas) for each document.

We used a python package **tethne** to generate output files that can be used to export the topics' evolution over the timespan, the topics' most common words into simple txt files.

DTM Parameters:
$dtm-darwin64 ./main --ntopics=3
--mode=fit --rng_seed=0 --initialize_lda=true --corpus_prefix=data/WWW/
--outname=data/WWW/output
--top_chain_var=0.9 --alpha=0.01
--lda_sequence_min_iter=6
--lda_sequence_max_iter=20
--lda_max_em_iter=20

EXPERT IDENTIFICATION USING SNA

We constructed several citation/collaboration networks using custom developed python scripts namely
1.) Direct citation network: where a directed edge (i,j) represents a citation from paper i to j.
2.) Co-citation network: Nodes are articles and edges are added when 2 articles are cited together indicating they are similar works.
3.) Co-author network: is built based on which author writes papers with whom. Authors are nodes while edges represent co-authorship on a paper.

Fig 1: *Direct citation network after filtering 84M citation edges to nodes having at least 100 citations*

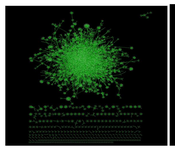


Fig 2: *Zoomed in view of largest connected component*

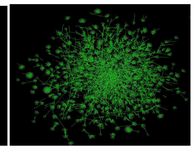


Fig 3,4: *Top nodes obtained after filtering by degree and betweenness centrality on direct citation network*

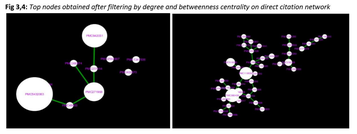


Fig 5: *Co-citation network constructed from 1M citations dataset and filtered by nodes with at least 100 citations.*

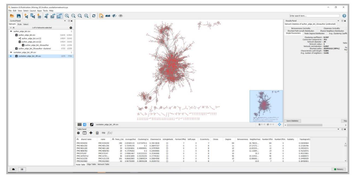


Fig 6: *Clusters identified after executing community detection algorithms on a co-citation network of 1M citations. Each cluster indicates a set of articles in a specific research topic*

Fig 7: *Zoomed in view of a cluster. Size of the node represents the number of times the articles was cited. Edge thickness represents the number of times the articles were co-cited.*

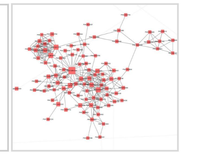


Fig 8:*Clusters identified after executing community detection algorithms on a co-author network of 1M authors.*

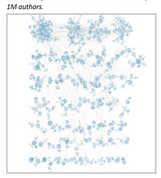
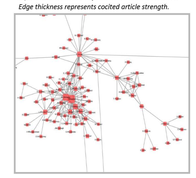
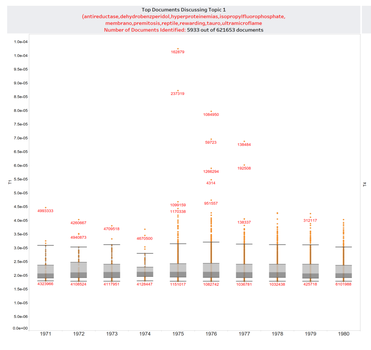


Fig 9 *Zoomed in view of a cluster. Size of the node represents the number of times the articles was cited. Edge thickness represents coicited article strength.*



METHODOLOGY
- All the data was downloaded and imported into the postgresql database. The outputs from the python scripts consisted of node and edge files with node attributes such as number of citations per articles or number of publications and edge attributes such as number of co-citations/number of coauthored publications.
- These files were imported into the Cytoscape tool which is an open source software platform for complex network analysis and visualization. The networks were further filtered using the node attributes(times_cited and number of publications per author) to identify the key players/authors.
- We also executed community detection algorithms[3](Glay) on the co-citation and co-authorship networks to identify key groups/topics and identified experts in those cluster groups through page rank and several centrality metrics.

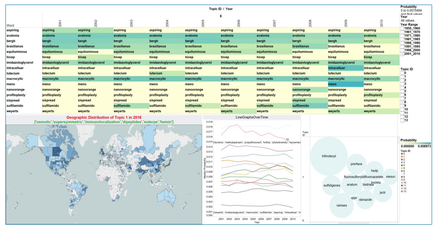


REFERENCES & FUTURE WORK

CONCLUSION

We have developed an interactive **visual–analysis** tool that enables a general user/ subject matter expert to understand the key topics that have been researched over the previous 50 years and organize the documents accordingly using an unsupervised bayesian probabilistic model known as Dynamic Topic Model based on LDA (Linear Dirichlet Allocation).

Interactive Graphic User Interface (GUI)



FUTURE WORK

- Experimentation with parallel versions of DTM for faster training of the model using abstracts as opposed to titles(for years > 2010)
- Using a memory fat machine to compute the collaboration networks for the entire datasets instead of subsets.
- Implement algorithms to help clean-up affiliation address data that will allow better visualization of geographic distributions.
- Include context information within the topic models such authorship information or geographical coordinates associated with documents.

REFERENCES

1.Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.
2.Blei, David M., and John D. Lafferty. "Dynamic topic models." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.
3.Gang Su, Allan Kuchinsky, John H. Morris, David J. States, Fan Meng; GLay: community structure analysis of biological networks. Bioinformatics 2010; 26 (24): 3135-3137. doi:10.1093/bioinformatics/btq596
4.http://socialcomputing.ing.puc.cl/uploads/DynamicTopicModellingTutorial.pdf
5.https://www.cognizant.com/whitepapers/identifying-key-opinion-leaders-using-social-network-an alysis-codex1234.pdf
6.http://nealcaren.web.unc.edu/a-sociology-citation-network/
7.https://github.com/magsilva/dtm
8.http://web.mit.edu/seyda/www/Papers/ECIR_extended.pdf
9.Newman,MarkEJ."Coauthorship networks and patterns of scientific collaboration."Proceedings of the national academy of sciences101.suppl 1 (2004): 5200-5205