An Author Topic Analysis on NCI DCP/DCCPS PIs

Dingcheng Li
BSI
Mayo Cinic
Rochester, MN
li.dingcheng@mayo.edu

Janet Okamoto
Hemotology/Oncology
Mayo Clinic
Scottsdale, Arizona
okamoto.janet@mayo.edu

Hongfang Liu*
BSI
Mayo Clinic
Rochester, MN
liu.hongfang@mayo.edu

Scott Leischow
Hemotology/Oncology
Mayo Clinic
Scottsdale, Arizona
leischow.scott@mayo.edu

ABSTRACT

To facilitate the cancer study at Mayo Clinic, Mayo CPC (Cancer Population Control) initiated a study on the landscape of NCI funded principle investigators (PIs), who focus on cancer preventions, cancer control and population science. In this work, we conducted a bibliometric analysis on such research by applying author topic modeling (ATM) on MEDLINE citations published by currently-funded PIs from both DCP (Division of Cancer Preventions) and DCCPS (Division of Cancer Control and Population Science. Our initial results show that ATM can address the issue of research interests reasonably. Furthermore, a network involving authors, topics and words can be established for more detailed bibliometric analysis. This network may also be useful to grantees and funding administrators in suggesting potential collaborators.

1. INTRODUCTION

The Division of Cancer Prevention (DCP) of the National Cancer Institute (NCI) supports research to determine and reduce a person's risk of developing cancer, as well as research to develop and evaluate cancer screening procedures while the Division of Cancer Control and Population Science (DCCPS) of NCI supports a comprehensive program of genetic, epidemiologic, behavioral, social, and surveillance cancer research. Large number of productive research has been done thanks to those supports. One question we may ask is how good those research projects are. There are diverse approaches to make such assessments: annual research reports, estimation of research publications with the number of publications or the prestige of the journals where the publications go and how many research discoveries or proposed methodologies have been translated into clinical practices. However, up to now, there has not been a full quality assessment being done based on bibliometric analysis so that the full landscape of those studies is not crystal clear yet.

Mayo Cancer Center's Cancer Prevention and Control (CC-CPC) attempts to integrate the research efforts for quality improvements and meanwhile makes clear the future directions of Mayo CPC development. Therefore, a study is initiated on the landscape of NCI funded principle investigators (PIs), who focus on cancer preventions, cancer control and population science.

In this paper, we apply ATM [1] to simultaneously model the content of documents and the interests of authors. Namely, given the broader NCI DCP/DCCPS research field, we attempt to discover topics as well as general research interests utilizing MEDLINE citations for currently funded NCI DCP/DCCPS investigators. Meanwhile, we attempt to discover the hidden connections between the two groups of studies.

2. MATERIALS AND WORKFLOW

All documents used in this study are limited to the abstracts of both PIs of DCP and DCCPS published after 2008 when the recent grants are awarded. In order to better deploy the citations, we downloaded all MedLINE abstracts and indexed them with Lucene.Then, we extracted all abstracts published by all PIs (614 PIs for DCP and 809 for DCCPS) via Lucene JAVA API with PI name as the search field. Considering the duplications of author names, we used author affiliations as the main disambiguation fields to find unique authors. Since in this study, we are only concerned about all PIs, we just assume that each article was written by one author. The document set includes those MEDLINE citations with abstract available, resulting in 9538 abstracts for DCP researchers and 8264 abstracts for DCCPS researchers.

For each document, we remove stop words and filter words based on Term Frequency-Inverse Document Frequency (TF-IDF). Namely words with high document frequencies and relatively insignificant for single document are removed.

We ran the ATM developed by [2] on it for 200 iterations. Topic number T is selected as 20 (this is determined by held out perplexity). The hyperparameters α and β are fixed as 50/T and 0.01 respectively.

3. RESULTS AND ANALYSIS

3.1 Topic Proportions

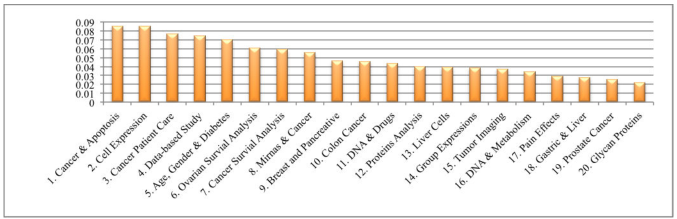

Figure 1: Topic Proportion of DCP Publications

Figure 1 and Figure 2 show the ordered proportion of the 20 topics for DCP publications and for DCCP publications respectively. In order to find out what each topic is focused on, we assigned each topic a name based on the top 20 words and also assigned a number to refer it. For DCP publications, most of the 20 topics involve specific cancer preventions while the top five focus on studying cancer mechanisms from genomic source. It looks that latest cancer studies attempt to understand the internal causes of pathological changes from biological structures. For DCCP publications, there are some overlapping topics, such as breast cancer,

*Dr. Liu and Dr. Leischow are both senior authors with equal contributions

BCB '14, September 20 - 23 2014, Newport Beach, CA, USA
ACM 978-1-4503-2894-4/14/09.
http://dx.doi.org/10.1145/2649387.2660798

colorectal cancer and ovarian cancer. But obviously, there are not so many specific cancers as discussed in DCP. Another evident trend is that diets, intervention treatment, survival or tobacco control, (which are related to cancer control) and gender, age, data or network analysis or statistical modeling, (which are related to population science) are involved. In the following section, we will look into the details of each topic to get a more fine level of understanding what key words each topic is composed of.

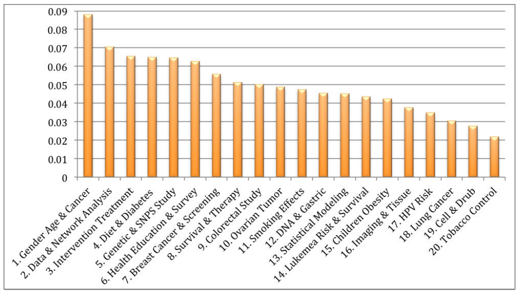

Figure 2: Topic Proportion of DCCPS Publications

3.2 Author Topic Relation

Now, if we turn to the correspondence of top authors (if a PI has more than 0.01 portion of articles in one topic, he or she would be counted as a top author) and topics in Figure 3 and Figure 4, more interesting patterns can be found.

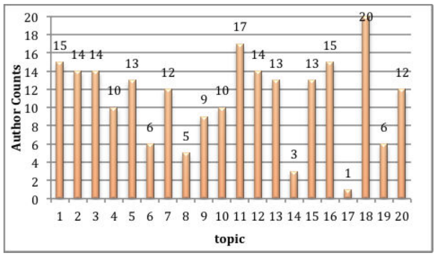

Figure 3: Author counts on topic maximum of DCP

Figure 3 shows that the topic with highest ratio is Gender, Age & Cancer. This means that there are 20 PIs whose highest part of research is on Gastric and Liver Cancer and the topic, Cancer Patient Care enjoys highest concentration since 20 PIs' research ratio on it is above 0.01. Yet, there is 0.09 of the whole cancer prevention study is put on the topic, Gender, Age & cancer. The number of PIs on this topic is also large, 17 PIs has more than 0.01 portion of research on it and 15 PIs has highest research ratio on it though it is not the highest in both counts. We can also see that one PI's highest research proportion is in topic 17, HPV risk although there are three PIs' research proportions on it, which is higher than 0.01. For topic 20, Glycan Proteins, only 3 PIs has more than 0.01 research ratio on it while 12 PIs' research focus on it. This shows that glucose seems to be related to many cancers though it does not play a decisive role. Meanwhile, this also implies that some new investigators may start to devote their research to this topic.

Similar contrast can be seen from Figure 4 for DCCPS grants. The topic, Cancer & Apoptosis has the highest proportion, up to 0.085 and about 19 PIs have more than 0.01 proportion of research on it. Yet, only 10 PIs have highest proportion on it. Similarly, the topic 20, Tobacco Control, which has only 3 PIs' research ratio higher than 0.01 on it while there are 20 PIs, who mainly focus on this topic. This may be related to the implementation of the Family Smoking Prevention and Tobacco Control Act of 2009. Since that on, more and more investigators started to oversee tobacco regulation activities. This also seems to show that although tobacco control is not the most important part in cancer prevention study, it is an indispensable part of cancer prevention research because tobacco is relevant to many cancers. Another interesting thing is to look at co-occurrence of authors among multiple topics (for simplicity, we only consider two). It can reflect two aspects, one on the closeness of two topics (the two or more can be subtopics of a big topic) and the other on interactions of two topics (they may not be related but depend on each other). It is found that T15 and T8 co-occur together 10 times, ranking the highest. It means that 10 authors study both topics. Both topics involve genetic expressions, cell, and protein. The combination of T16 and T8 follows closely where topic 16 is about lung tumor study from gene and cell level. The topic dependence relation can be illustrated by the large number of topics co-occurring with T2 (intervention).

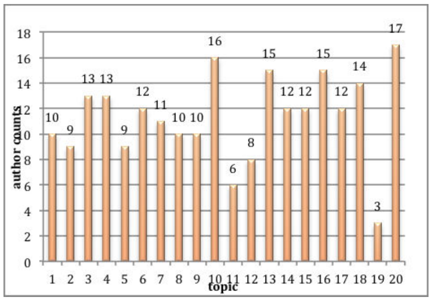

Figure 4: Author counts on topic maximum of DCCPS

4. DISCUSSION AND CONCLUSION

In this work, we employ ATM to model principle investigators on DCP and DCCPS granted researchers and their topics of research based on PubMED literatures.

The results show that this approach can efficiently cluster collections of articles into discriminative categories without any supervision. It can associate topics to authors in a high accuracy. This indicates that ATM can be utilized to infer the identity of an author of articles using topics generated by the model. The relevance of this analysis to DCP and DCCPS researchers is at least twofold. First, this analysis is a "proof of concept" that can be beneficial assess the change over time in cancer study as new projects are funded and collaborative science in this area changes. The results can thus be used to assess the extent to which new research reflects the funding priorities of the two organizations. Second, ATM outcomes can be used by investigators to assess who is conducting research in a particular research domain in order to foster collaborative science. By fostering collaborative science in cancer study, it becomes possible to speed advances in that science by fostering communication between scientists that can avoiding un-needed duplication and impact decision-making on new science that can benefit regulatory decision-making.

Acknowledgments

This study was made possible by National Science Foundation ABI:0845523, National Institute of Health R01LM009959A1 and R01GM102283A1.

5. REFERENCES

[1] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[2] M. Steyvers and T. Griffiths. Matlab topic modeling toolbox 1.4. *URL http://psiexp. ss. uci. edu/research/programs_data/toolbox. htm*, 2011.