A Bibliometric Analysis on Cancer Population Science with Topic Modeling
Ding-Cheng Li, PhD¹, Majid Rastegar-Mojadar, MS¹, Janet Okamoto, PhD², Hongfang Liu, PhD¹, Scott Leichow, PhD²
¹Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN;
²Health Science Research, Mayo Clinic, Scottsdale, AZ

Abstract

Bibliometric analysis is a research method used in library and information science to evaluate research performance. It applies quantitative and statistical analyses to describe patterns observed in a set of publications and can help identify previous, current, and future research trends or focus. To better guide our institutional strategic plan in cancer population science, we conducted bibliometric analysis on publications of investigators currently funded by either Division of Cancer Preventions (DCP) or Division of Cancer Control and Population Science (DCCPS) at National Cancer Institute. We applied two topic modeling techniques: author topic modeling (AT) and dynamic topic modeling (DTM). Our initial results show that AT can address reasonably the issues related to investigators’ research interests, research topic distributions and popularities. In compensation, DTM can address the evolving trend of each topic by displaying the proportion changes of key words, which is consistent with the changes of MeSH headings.

Introduction

Division of Cancer Prevention (DCP) and Division of Cancer Control and Population Science (DCCPS) of National Cancer Institute (NCI) are two main divisions of supporting research in cancer prevention and control. The former’s support focuses more on research to determine and reducing a person’s risk of developing cancer as well as research to develop and evaluate cancer screening procedures while the latter supports a comprehensive program of genetic, epidemiologic, behavioral, social, and surveillance cancer research. The major scientific collaborations and research networks for both DCP and DCCPS are spreading at more than 100 sites across the United States and the grants involve investigator-initiated grants; postdoctoral training; and specialized resources for researchers. Many productive studies have been published thanks to those supports. However, whether those publications are consistent with the missions of the two divisions, how “hot topics” fostered by principle investigators (PIs) supported by the two divisions are evolving and how to identify future research directions from those publications are not fully clear yet. Understanding of such aspects offers the potential to foster better collaborative science in cancer prevention and controls, provide landscapes to both divisions to make wiser decisions on grants allocations and assist institutional strategic plan in developing cancer population science.

An intuitive approach to address those aspects is to conduct bibliometric analysis on publications of corresponding principle investigators supported by the both divisions. Bibliometric analysis applies quantitative and statistical analyses to describe patterns observed in the contents and citations among a set of publications and has been widely used in library and information science to evaluate research performance and identify previous, current, and future research trends. Diverse methodologies, such as graph construction and modeling [1], latent semantic indexing (LSI) [2] or latent Dirichlet allocations (LDA) [3] have been utilized for this purpose. Co-citation analysis is an important subset of bibliometrics and is subdivided into author co-citation, journal co-citation, keyword co-citation and so on. Keyword co-citation analysis and the author-citation analysis are often intertwined. Jeong and colleagues proposed content-based author cocitation analysis [4]. They extended author co-citation analysis methods by incorporating citing sentence similarity into citation counts. They used citing sentences to obtain the topic relatedness between the cited authors and the cited sentence. This similarity is measured by topic relatedness between two citing sentences.

In this paper, we report the use of two topic modeling techniques: author topic modeling (AT) [5] and dynamic topic modeling (DTM) [6] to conduct a bibliometric analysis of publications of funded cancer population science researchers from those two divisions. Different from traditional approaches, AT and DTM integrate content analysis with

Figure 1 Workflow of the framework
author and temporal as random variables so that contexts, the author interests and topic evolutions are modeled simultaneously.

Besides, another advantage of our approach is that keyword co-citation and author co-citation are modeled together while at the document level rather than sentence level. Topics generated by those two models can be more consistent with author interests and dynamically evolves with time.

**Methods, Materials and Workflow**

The overall components and steps of the framework are illustrated in Figure 1. The workflow consists of six steps of information processing. Via Retrieve publications queried by PI, abstracts of both PIs of DCP and DCCPS are retrieved from PubMed search engine. Author affiliations are employed to remove duplications of author names (we ignore the few overlapping of PIs from both sources for now). Since in this study, we are only concerned about all PIs, we just assume that each article was written only by PIs and co-authors are ignored for now. The second step aims at preprocessing the data. For each document, we remove stop words and filter out words such as cancer, which appear in almost in each abstract, based on Term Frequency-Inverse Document Frequency (TF-IDF).

On the third step, three sparse matrices for both DCP and DCCPS datasets are constructed respectively. The first matrix is the author-document matrix (short for AD) where the row is the document and the column is the author names (PI in our work, in fact). The second matrix is the document-word matrix (DW) where the row is the word and the column is the document. The third matrix is the document-word matrix by year (YDW). Both AD and DW are used as input for AT in the fourth step respectively. As an extension of LDA, AT incorporates authors by adding one more variable, which is uniformly assigned by a set of authors, an observed set in some corpus. Hence, in AT, a topic is chosen from a distribution over topics specific to that author, and then the word is generated from the chosen topic. This model can be understood as a two-stage stochastic process. An author is represented by a probability distribution over topics, and each topic is represented as probability distributions over words.

Similarly, in the fifth step, YDW and corresponding temporal information will be used as input for DTM. As another extension of LDA, DTM takes the temporal changes of topics into considerations. Therefore, the document is now sequentially organized. That is how YDW is grouped. DTM can be understood as hidden Markov models (HMM) version of LDA. But different from HMM, the temporal dependences occur to the both hyperparameters \( \alpha \) and \( \beta \) rather than to the hidden states.

For the last step, we analyze the temporal changes of MeSH terms. It aims at making a comparison with topic changes generated by DTM.

**Results and Analysis**

There are 614 PIs for DCP and 809 for DCCPS. The document set includes those MEDLINE citations with abstract available, resulting in 9538 abstracts for DCP researchers and 8264 abstracts for DCCPS researchers in Step 1. We ran the AT developed by Steyvers et al [7] on both the datasets of DCP and DCCPS for 200 iterations. Topic number \( T \) is selected as 20 based on the perplexity test. The hyperparameters \( \alpha \) and \( \beta \) are fixed as 50/T and 0.01 respectively following empirical selections of previous research. The DTM package we utilize is developed by Blei et al [6]. Topic number \( T \), the initial hyperparameter \( \alpha \) and \( \beta \) are identical to those of AT.
Topic Proportions

Figure 2 and Figure 3 show the ordered proportion of the 20 topics for DCP publications and for DCCPS publications respectively. In order to find out is defined as the number what each topic was focused on, we assigned each topic a name based on the words with posterior probabilities higher than some threshold and also assigned a number to refer it. For DCP publications, most of the 20 topics involved specific cancer preventions while the top five focused on studying cancer mechanisms from genomic source. It looks that modern cancer studies attempt to understand the internal causes of pathological changes from biological structures. Topics of both DCCPS and DCP publications involve breast cancer, colorectal cancer and ovarian cancer. But obviously, there are not so many specific cancer studies in DCCPS as in DCP (more than half). It is understandable since research of DCP aims at solving problems of cancer preventions from disease itself while that of DCCPS at study of characteristics of cancer population and epidemiological control of cancer spreading. Thus, topics related to statistical studies (Topic 1, 16, 13 and 20 for example) can be found in DCCPS. In the following section, we will look into the details of each topic to get a more fine level of understanding the relationship between author, topic and key words.

Author-topic Relation Network Analysis

Figure 4 and Figure 5 are the author-topic relation network constructed based on the AT outputs with an open source visualization package called Cytoscape [8]. The networks are composed of three types of nodes—red, blue and green, representing the topic, the PI and the key word respectively.

In network analysis, edge-betweenness (EBC) is one of the main measures to tell how important/popular a node is. It is defined as the ratio of shorted paths from all vertices to all others that pass through that node (the red float numbers in the figures). In the graph, the higher the EBC is, the larger the size of the node is and the more edges the node involves.

For example, if we want to see how popular a topic is, we can count the number of edges from a topic node to PI nodes it connects. Similarly, we can also see how many topics each PI has been working on by the number of edges starting from that PI to the topic node and similarly how many key words each topic included.

Due to space limitation, we cannot show all details of the network. Yet, a few interesting patterns can be found. Firstly, the topic with the high ratio was not necessarily the one involving the most number of PIs. For example, the EBC of Topic 1 of DCP and DCCPS (Cancer Apoptosis and Gender, Age & Cancer) are both 0.087 and rank in the lower level compared with 0.012 for Topic 3 (Cancer Patient Care) of DCP and Topic 15 (Children Obesity) of DCCPS, which rank first respectively. This seems to show that the popularity of a topic is not related to its proportion. It may be related to the fact that research in some topics can easily yield publications.

Secondly, the top 10 key words of DCP and DCCPS (the green nodes on the right) suggest that researchers in DCP focus more on cancer prevention and treatment while those in DCCPS more on cancer control and population science. For example, in DCP, treatment (which suggests treatment study) and protein or effects (which suggests prevention study) rank high in DCP while in DCCPS, risk or control (suggests cancer control) and women or age (suggests population) rank high.

Thirdly, some “star” PIs are clearly displayed from the network. The higher EBC a PI is, the more diverse his or her research is. A warning is that the ranking based on EBC for PIs does not mean that he or she is not productive if his or her EBC is low. It only means that he or she focuses more on some specific topics.
Two of the resulting topics (Glycan Proteins for DCP and Tobacco Control for DCCPS respectively) are illustrated in Figure 6. Each part is composed of three parts: (a) the top ten words from the inferred posterior distribution from 2008 to 2014, (b) the posterior estimate of the frequency as a function of year of several words from the same topics as in (a), (c) example articles throughout the collections of PIs’ publications of both DCP and DCCPS, which exhibit these topics. The reason that we selected those two topics was that both topics were not high in topic ratios while ranked high considering the number of PIs involved. The plots were scaled to give an idea of the shape of the evolving trend of the words’ posterior probability. From the two plots and the ranking changes of those top ten words, it seems that we can say that there were some minor changes (or even some fluctuations, like what the word bitterness shows) for the research focuses of DCP PIs’ on glycan study while the research trend of tobacco control seems to be more stably growing. 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 and do more close research on the relations between tobacco control and cancer population.

In order to validate what we illustrated above, we made a sample statistical analysis on the yearly changes of MeSH terms as well. For example, for DCP, glycaemia-related as a MeSH term is hypoglycaemia and hyperglycaemia. It is found that both of them were used in a similar frequency as normoglycaemia (a little bit decreasing trend). The corresponding MeSH term of bitterness, namely pain, has been used in a slightly increasing trend. The word, malondialdehyde, which can be found in the MeSH list, showed decreasing trend from 2009 to 2012 (but not in 2008 and 2013, 2014). Similarly, for MeSH terms in DCCPS, both tumor and cardiopulmonary had a slightly increasing trend and fitness the contrary. Hence, as illustrated and validated, DTM captures the main themes of the two topics and can be used to inspect trends of word usage within them.

**Discussion and Conclusion**

In this work, we employed AT to model principle investigators funded through NCI’s DCP and DCCPS divisions and their topics of research based on PubMed literatures. We also applied DTM to detect how related research topics evolved over time from 2008 to now. Both modeling techniques have been shown to be an effective approach in modeling documents from computer sciences, tobacco regulatory science [9] as well as more general fields, like publically available emails, collections of diverse research articles. No research has been done in modeling a constraint domain like cancer prevention and control. The results show that this approach can efficiently cluster
collections of articles into discriminative categories without any supervision. More interestingly, AT can associate
topics to authors and DTM can reflect topic changes with high accuracy. The relevance of this analysis to DCP and
DCCPS researchers is at least twofold. First, outcomes of AT can be used by investigators to assess who are
conducting research in a particular research domain in order to foster collaborative science. By fostering
collaborations in cancer studies, and thus fostering communication between scientists, it becomes possible to speed
advances in the field that can hopefully prevent duplication and impact decision-making on new streams of scientific
inquiry. Second, this analysis is a ‘proof of concept’ that can be beneficial to assess the change over time in cancer
prevention and control, 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.

Limitations and Future Work

One limitation for this approach is that AT assumes that the topic distribution of each word in one document is only
associated with one of the known authors, thus correlations of authors cannot be reflected from words of the same
document and instead, must be found across multiple documents, which have the same authors. Nonetheless, this
limitation can be overcome if we introduce the topic-author associations as multiple to multiple. Hence, in our future
work, we will extend AT into group AT. In addition, considering that research interests of PIs are, more often than
not, fluid and dynamic rather than static, it would be reasonable to model temporal changes over authors as well.
Therefore, it would be ideal to combine author-topic modeling and DTM together so that we can model the
evolution of research topics and authors’ research interests in one model. One natural extension would be to build
predictive models where we can assign authors to unknown articles or we can predict how the research interests of a
PI may change in the future. More reasonable parameter estimation should be made accordingly. Another limitation
is that the topic name may not fully and accurately reflect the meaning of the topic, though generally naming is
based on top words of topic distributions. In order to overcome this, in our future work, we are taking the approach
of maximizing mutual information between a label and a topic model proposed by Mei [10] to estimate the fitness of
topic labels. Beyond, as noticed, top topics by proportion do not correspond to the top by author connectedness
count. We did not explore deep why and how this happens. In future work, we will integrate with journal impact
factors or number of publications per author so that we can distinguish broadly popular topics written about by many
authors and specialized topics that only a few authors write a lot about.

Acknowledgment

This study was made possible by National Science Foundation ABI:0845523, National Institute of Health
R01LM009959A1 and R01GM102283A

References

[1] J. Whittaker, Graphical models in applied multivariate statistics: Wiley Publishing, 2009.
[2] S. T. Dumais, "Latent semantic analysis," Annual Review of Information Science and Technology, vol. 38,
pp. 188-230, 2005.
[3] K. Lu and D. Wolfram, "Measuring author research relatedness: A comparison of word-based, topic-based,
and author cocitation approaches," Journal of the American Society for Information Science and
Technology, vol. 63, pp. 1973-1986, 2012.
[4] Y. K. Jeong, M. Song, and Y. Ding, "Content-based author co-citation analysis," Journal of Informetrics,
vol. 8, pp. 197-211, 2014.
[5] 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, 2004, pp. 487-
494.
[6] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proceedings of the 23rd international
conference on Machine learning, 2006, pp. 113-120.
[7] T. G. Mark Steyvers. (2014, Oct. 7). Matlab Topic Modeling Toolbox 1.4. Available:
http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
[8] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, et al., "Cytoscape: a software
environment for integrated models of biomolecular interaction networks," Genome research, vol. 13, pp.
2498-2504, 2003.
[9] D. Li, J Okamoto, H Lii, S Leischow "An Author Topic Analysis of Tobacco Regulation Investigators," in
The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Tai Wan, 2014.
[10] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in Proceedings of the
13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 490-499.