COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Alan Porter (Search Technology, Inc. and Georgia Institute of Technology), January 2022

Title: Exploring Causes and Cures for COVID-19 through Improved Access to Biomedical Research

Alan Porter CIC Profile

NSF Award #: 2029673

YouTube Recording with Slides

January 2022 CIC Webinar Information

Transcript Editor: Saanya Subasinghe

Transcript

Alan Porter:

Slide 1
Thanks everyone, I wanted to share some results from an NSF RAPID project that we did. With the title here - and it's basically mining the biomedical literature in ways to try to get access conveniently and some different perspectives.

Slide 2
Here's my background. I'll not go through that. Basically we're technology watchers - trying to gather data and extract some usable knowledge from it.

Slide 3
A particular project here was exploring causes and cures for COVID through improved access to the research and I want to just touch on three elements. One is [what] we're calling intelligent bibliometrics to try to extract information from these abstract texts. Second is a peek at a recommender system to do some literature based discovery beyond the data set we're actually analyzing. And third is a dashboard to share results.

Slide 4

Our shorthand for this is 'tech mining'. Data at hand, here, are using the main National Library of Medicine's, kind of, core COVID search strategy. We used this straight through for a year and a half. And I'll be talking about two different subsets. One is about 60,000 records through early October 2020 on which we did a lot of the intelligent bibliometrics type work and the other is the most recent version of the dashboard from January 1st - to show you what it looks like right now.

Slide 5

Okay, element number one: the tech mining is text analyses of science technology and innovation data resources, usually in the form of abstract records of research publication or PAT, most typically. Some of the things we've done here - one is look within the body of that research for what topics show accelerating attention and try to do that in a timely fashion. I won't be talking about that today. Second is tracking topics over time.

Slide 6

Here's the peek at that led by our colleague Yi Zhong in University of Technology, Sydney. One draws out by clustering the term usage or what are some topical clusters of interest over time?

Slide 7

The process is, in essence, take one time period group the topics together based on co-occurrence of terms, and then do this in successive periods, and then look at links amongst those topics. And I've got a couple of references here i won't labor but they're tagged on at the back of the presentation.

Slide 8

Here's a zoom in on a little piece of it. Our latest analyses back in 2020 showed rapid spread was a topic of interest. And that had forbears in very recent treatment of infectious diseases- a little earlier to studies noting time. And then going back into some of the previous coronaviruses - some work with PCR back in 2006 concentration. Tracing back to SARS back in 2003. So one can do some links over topics and over time.

Slide 9

Our second element - the literature-based discovery - and here we've used it just for a very simple recommender system. See if there's a topic you are interested in within the COVID literature. Might there be some useful research to you outside that literature?

Slide 10

And the process here - we generate a dataset just as a simple illustration out of about 30 or more topics co-morbidities with one that we just picked here. And then within the COVID literature, in this case that's 60,000 records, looking for the documents most related to that. Then do some text analyses there to clean the literature that detects that, and looking for what terms have especially high frequency in those documents, and conversely especially low frequency, calling that a knowledge model. Then going out and looking at calculations of the firm frequency [and] inverse document frequency for those terms

in the full PubMed Medline dataset. Those aren't identical [inaudible]. And to identify abstracts of articles that might be of interest to somebody pursuing a particular topic, like co-morbidity.

Slide 11
Here's a little illustration of three of the topics from comorbidities where we've added in a citation count as a second criteria. So if you're high on being cited and you're using the terms similar to the co-morbidity usage within COVID literature, [this] might be of interest to you. Maybe take a look at these and it could be three articles, or we could go all the way down to loadings on the 33 million, which I don't think anybody would want.

Slide 12 - https://sites.google.com/searchtech.com/covidproject/home
The last element I wanted to go through was taking a look at our dashboard build-up. And here is the project dashboard and within that I'm just mentioning a vaccine bullseye that our colleagues at BizInt put together, which is interesting to track over time. I'm going to take a look at our PubMed dashboard. And that's now spotlighting the 153 - 150,000 abstract records as of January 1st. And there is a demo Fthere, a couple minutes, to explain how to use it. Here's some of the different fields of data available. And let me just do a quick pop-in here's one version where we can go in and we could spot something of interest and I'm just gonna - try to spotlight. And I think we're getting some sluggish response, here. Research in Iran - we might go through that and see what some folks over there are doing with respect to co-morbidities. And we might even spot an article somewhere in here that draws our interest. And here we can pop up that abstract record. So to do things quickly, there's just one other illustration. If we wanted to spotlight the most recent work and that's from December of last 2021. This is incomplete because coding, categorizing still is going on, cataloging, and so on, but we could go in here again and pick what might be of interest. [Freezing video] At Columbia - and there's some of the 46 articles from - associated with an author from Columbia in the last month. So that's our demo- if anybody wants to go to the dashboard we'd love it. It's still a work in progress.

Slide 13
And I'm basically done. There we go - Resources and Finish. Thanks for your attention!