Methodological and Technical Challenges in Big Scientometric Data Analytics

Jian Qin, Syracuse University
Mark Costa, Syracuse University
Jun Wang, Syracuse University

Abstract
Scientometric analytics faces new technical and methodological challenges in using large-scale metadata as the source of its analysis. This poster reports such challenges encountered in a scientometric analytics project that uses the metadata in GenBank, as well as the implications for data quality, processing, and analysis in scientometric analytics and for metadata management.
Keywords: Scientometric data analytics; Data readiness; Data repositories; Data quality
Citation: Qin, J., Costa, M., Wang, J. (2015). Methodological and Technical Challenges in Big Scientometric Data Analytics. In *iConference 2015 Proceedings*.

Acknowledgements: This research is sponsored by the NSF's Science of Science Policy Program, grant number 1262535.
Contact: jqin@syr.edu.

1 Introduction

Scientometrics is a field that studies the patterns, trends, and dynamics of science as an enterprise by using quantitative data and methods. Research has been published from investigating productivity, collaboration, and impact of individual scientists, institutions, and disciplinary fields, as well as mapping of the history and trends of science. Scientometric analytics is considered as a useful approach to understanding the science enterprise and providing support for science policymaking.

Traditionally, scientometric analytics is done by using publication metadata such as those in the Science Citation Index and Scopus databases, among others, as the data source. The increasingly fast pace in scientific data growth in the last few decades has brought about new challenges for scientometric analytics. Large-scale data repositories that contain massive amount of metadata are common in today's data-driven science: the Long Term Ecological Research Network (LTER) and the National Center for Biotechnology Information (NCBI) data repository GenBank are two examples. These data repositories contain metadata that not only describe datasets generated in research lifecycles but also have links to publications that are produced based on these data, which offer a great (big) data source for scientometric analytics at an unprecedented scale and complexity.

The authors of this poster have been working on a scientometric analytics project using the GenBank metadata (http://www.ncbi.nlm.nih.gov/genbank) that studies the structures and dynamics of research collaboration networks. This poster will focus on the methodological and technical challenges we encountered in this project and their implications for data quality, processing, and analysis in scientometric analytics as well as for metadata management.

2 Background of GenBank

GenBank is an international data repository hosted by the National Center for Biological Information (NCBI). It was established in the early 1990s (though ground work started in the late 1980s) and has grown to become the largest data archive for DNA and RNA sequences.

The GenBank FTP server stores semi-structured, compressed genetic sequence data files submitted by researchers from around the world. Most of these compressed files are anywhere from 2MB-80MB. As of 8/16/2012 when we downloaded the data, it contained 1,723 data files and 129 index files, all in the compressed .gz format. The size of the compressed sequence data was approximately 75GB and that of the index files 8GB. Each GenBank record contains both metadata describing the submission as well as the genetic sequencing information. The sequencing information comprises the bulk of the file size (over 90% of the file in many cases) and is not needed for the purposes of our study.

A GenBank data record is structured in three main blocks: metadata about a genetic sequence, features of the sequence, and the sequence data itself. There are three types of author data in the metadata block: publication authors, sequence authors, and patent authors. Table 1 shows a sample GenBank record's metadata block, which includes basic information about the sequence, related publications, and sequence submission date and author(s). The last reference is always for submission authors, affiliations, and date.

Table 1: A sample metadata record for direct submission and publications

LOCUS	FJ208946 1204 bp mRNA linear VRT 13-APR-2009
DEFINITION	Gillichthys mirabilis enolase 1 isoform b mRNA, partial cds.
ACCESSION	FJ208946
VERSION	FJ208946.1 GI:226441954
KEYWORDS	.
SOURCE	Gillichthys mirabilis (long-jawed mudsucker)
ORGANISM	Gillichthys mirabilis Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei; Acanthomorpha; Acanthopterygii; Percomorpha; Perciformes; Gobioidei; Gobiidae; Gobionellinae; Gillichthys.
REFERENCE	1 (bases 1 to 1204)
AUTHORS	Bucciarelli,G., Di Filippo,M., Costagliola,D., Alvarez-Valin,F., Bernardi,G.and Bernardi,G.
TITLE	Environmental genomics: a tale of two fishes
JOURNAL	Mol. Biol. Evol. 26 (6), 1235-1243 (2009)
PUBMED	19270014
REFERENCE	2 (bases 1 to 1204)
AUTHORS	Bucciarelli,G., Di Filippo,M., Costagliola,D., Alvarez-Valin,F., Bernardi,G.and Bernardi,G.
TITLE	Direct Submission
JOURNAL	Submitted (13-SEP-2008) Stazione Zoologica Anton Dohrn, Villa Comunale, Naples 80121, Italy

3 Challenges in Big Scientometric Data Analytics

3.1 Technical challenges

The primary technical challenges fall into two broad categories – storage and retrieval, and analysis. In both cases, we need to develop better frameworks for determining sufficient computational capacity to support teams whose members' resource use is highly variable and expensive to coordinate and control. This means keeping a balance between cost, peak need, and idle resources provisioned (wasted capacity). The data collected so far include four main categories: authors, publications, sequence data submissions, and organisms represented by the NCBI Taxonomy. Each of these domains involves a number of attributes and cross relationships with data in other categories. For example, author data contain attributes of name, affiliation, and country, with relationships to all three other categories. The storage and retrieval of data in the first three categories can be handled by relational databases with support for data consistency, though the retrieval and processing time may be slowed down drastically if the query is complicated (i.e., involving multiple tables and criteria). The organism data, however, require recursive search over tables, in order to, for example, find a community structure of scientists who have sequenced data on the animal kingdom, which involves recursively finding all interactions over nested sub-categories such as species, genus, family, etc. The issues in computing efficiency and response time can significantly affect the project progress as computational activities generate more data in a fast pace.

Conducting analyses can also be computationally expensive. Current open source versions of analysis packages, such as R, do not natively support parallel computing. There are a number of packages that will bring such functionality to the analytic environment, but the approaches to analysis need to be reworked to leverage the added resources.

In this project we developed workflows and methods for the collection, extraction, parsing, and cleaning of metadata from GenBank, with which we transformed the data from semi-structured into structured format for further cleansing, all done in computationally manageable time. However, the idiosyncrasies of patent data submissions in GenBank require a significant amount of reconfiguration of the workflows and methods we developed. This is because scientists often submit a collection of base pair sections of data to GenBank and associate those submissions with one or two references. When these base pair sections of genetic sequences are submitted, scientists appear to file a patent associated with each base pair they sequence. Consequently, the data set for references to patents contains 26 million rows, in comparison to the 1.35 million rows of data on references to direct submissions of sequence data and publications.

3.2 Methodological challenges

The very large-scale of scientometric analytics makes manual operations on data unrealistic and impractical. All steps in this project, ranging from data cleaning, verification, linking, to transformation, need to be done using computational methods. In addition, we use complex network theory and measures to generate quantitative features or properties for the collaboration networks of our interest.

While network science has its own set of concepts and is a theory in itself, it is not the best candidate for interpreting the quantitative results from a social and policy perspective. Methodologically, we need to have a workflow and procedures to ensure the data quality and proper management while clearly understand the roles of different theories in the various aspects of scientometric analytics using Big Data.

4 Conclusion

The technical and methodological challenges entered in our project are not unique in scientometric analytics; projects using very large-scale data, especially those that are not well structured, will run into the similar challenges we had. While these challenges require more attentions from the i-professionals and researchers, our experience will help bring up more questions for this community to discuss and explore.