Machine learning to identify retracted articles in science

Subject areas: database design & management, machine learning, pattern recognition, academic publishing, data analytics

Background

Scientific research is typically presented in the form of manuscripts published in scholarly journals. These manuscripts undergo comprehensive peer review — a detailed vetting process intended to scrutinize the work and ensure its validity. While peer review is thought to generate high quality science, articles published through this mechanism are from time to time retracted, whether by the journal itself, the authors, or the authors' institution. This can be for a variety of reasons, ranging from benign issues related to typography, to concerns about the methods used, to outright academic fraud.

Retracted publications can be difficult to track. Different publishers and journal indexing systems may have different mechanisms to deal with retractions, and to link retraction notices to the original articles. As such, the impact of retracted work can propagate well after it is withdrawn. **There is currently no widely accessible and user-friendly repository to track retracted articles, and analyze their impact.** However, a journalist-run organization called Retraction Watch monitors the published scientific literature for erroneous, fraudulent, or otherwise compromised reports. They compile these hand-curated retracted articles in a database that can be accessed online through individual queries.

Objective

- To develop a user friendly and computationally friendly version of the Retraction Watch database, in order to support further studies of retractions
- To merge the Retraction Watch database with larger scientific indexing databases such as Scopus and PubMed
- To use machine learning to identify predictors of retraction, and assign a retraction likelihood for other articles.

Resources

- The Retraction Watch database
- Scopus, a repository of academic publications
- Pubmed, which can be accessed directly, or through Entrez Programming Utilities (E-utilities) that provides server side tools for efficient query and retrieval.
- The Queen's-affiliated Centre for Advanced Computing provides computing and storage infrastructure to support academic work at Queen's, such as the project described herein.

Deliverables

The goal of this project is to develop a comprehensive, searchable data environment merging Retraction Watch data with Scopus and/or PubMed, *using the retracted articles in the Retraction Watch database as a labeled dataset for machine learning classification tasks*.

Contacts:

Dr. David Maslove, Dept of Critical Care (david.maslove@queensu.ca)
www.conduitlab.org