International Journal of Population Data Science

Journal Website: www.ijpds.org

Automatic detection of cohorts in the haystack of academic publications

Tingay, K[1] and Anastasiou, A[2]

[1]Swansea University

[2]bibInsight, Swansea University Medical School

Introduction

Discovering suitable datasets is an important part of health research, particularly for projects using linked and/or cohort data, but with the proliferation of so many national and international initiatives, it is becoming increasingly difficult for research teams to locate real world datasets that are most relevant to their project objectives.

Objectives and Approach

To assist researchers in this, a bibliographic data analysis platform, bibInsight, is being developed which aims to identify potentially useful datasets, among other information, from large numbers of publications.

A search dataset on the broad topic of "dementia" was obtained from PubMed and abstracts mentioning cohort names were identified. These helped in determining the context within which cohort data are mentioned. Terms were also informed by specialist knowledge from the European Medical Information Framework (EMIF) Alzheimer's Disease project, and were used to train a one-class classifier that identified their structure.

Results

The initial PubMed search resulted in 129,040 articles, with 1961 identified cohorts. From these, the classifier was able to identify 1,588 as containing cohort information, and 373 as not, a precision of approximately 80%.

The terms identified by the classifier included "annual", "longitudinal", and "prospective", suggesting re-contact.

Further analysis of these 1,588 abstracts can find those most relevant to specific topics, co-authorship collaborations (suggesting data sharing agreements are already in place), and mention of linkage to other datasets. These can be done through keyword extraction, longitudinal data mining methods, linking author affiliations to institutions through the Global Resource Information Database (GRID), and graph databases, all of which have been built into the bibInsight platform.

Conclusion/Implications

'This approach shows promise in facilitating research by making literature searches more efficient. A key issue is working with unstructured text data which requires additional standardisation for analysis. Data reuse is a growing area, and publications can assist this, but better reporting would be beneficial.

Supported by EMIF grant 115372

http://dx.doi.org/10.23889/ijpds.v3i4.960