Analytics of Research Literature Quality through Content and Semantics of the Scholarly Literature

Fionn Murtagh*
De Montfort University, Leicester, UK - fmurtagh@acm.org

Arlene Casey
De Montfort University, Leicester, UK - arlenejanecasey@hotmail.co.uk

Samad Ahmadi
De Montfort University, Leicester, UK - sahmadi@dmu.ac.uk

Pedro Contreras
Thinking Safe Ltd., Egham, UK - pedro.contreras@acm.org

For textual content obtained from the scholarly, archival literature, we use open source, scalable NOSQL software including Apache/Lucene SOLR to support best match and some degree of similarity-based searching. We seek a scalable approach to addressing the challenge of determining semantics (or content) based quality assessment. Among other methods we can analyze the semantics of this data using the Correspondence Analysis platform. To handle large and very high dimensional semantic spaces, it is known from their symmetries (in particular, hierarchical symmetries) that such spaces are very simple in their structures, which both facilitates understanding the data and motivates the use of particular analytical methodologies such as hierarchical clustering.

Our longer term objectives are the following: use the scholarly literature, preprint servers, and all forms of reporting on research work and outputs; include also funding proposals since such documents constitute reports on (planned) research work and outputs; determine and track narratives of research disciplines and research areas. Medium-term, we define a document unit as a published article, or the main research funding proposal text. Using intra- and inter-document unit analysis, we will analyze the narratives that underly the content.

Keywords: big data analytics; research evaluation; latent semantics; information geometry and topology.