Personalized Feed/Query-formulation, Predictive Impact, and Ranking

Alex D. Wade[1][0000-0002-9366-1507] and Ivana Williams[1]

[1] Chan Zuckerberg Initiative, Redwood City CA, USA
awade@chanzuckerberg.com

Abstract. The Meta discovery system is designed to aid biomedical researchers in keeping up to date on the most recent and most impactful research publications and preprints via personalized feeds and search. The service generates feeds of recent papers that are specific and relevant to each user's scientific interests by leveraging state of the art embeddings and clustering techniques. Meta also calculates article-level *inferred* Eigenfactor® scores which are used to rank the papers. This paper discusses Meta's approach to query formulation and ranking to improve retrieval of recently published, and yet un-cited academic publications.

Keywords: Knowledge Graphs, Personalization, Article Ranking, Bibliometrics, Citation Networks, Scholarly Communication, Embeddings.

1 Introduction

The Meta discovery system [1] is a biomedical paper discovery and current awareness service developed by the Chan Zuckerberg Initiative. Meta has been designed to aid the biomedical research community in keeping up to date on the latest and most important research publications and preprints via personalized feeds and search. The Meta **Knowledge Graph** is a knowledge graph constructed from the biomedical literature, including coverage of PubMed and preprints from bioRxiv. Nodes in the graph correspond to entities such as authors, affiliations, papers, diseases, genes, proteins, MeSH terms, journals, etc. Edges connect two nodes based on various criteria. For example, an edge exists between two authors if they have co-authored a paper or an edge exists between two papers if one cites the other.

2 Personalized Feed Construction

A key goal of Meta is to provide a personalized and relevant experience, highlighting the most impactful recent papers of interest to the user. A challenge in creating this experience is to estimate or predict an individual user's research interests in order to create a personalized experience. An academic researcher's work is generally encapsulated within their publication history but might also be derived from their library of saved publications as well as through searches and other user interactions. Using both sentence and paper-level embeddings [2] [3] as well as unsupervised hierarchical clustering [4], Meta can algorithmically generate and score a set of queries that can serve as new feed query definitions, as well as matching a user with existing feeds. These feeds are intended to provide an initial experience which can then be further refined by the user to meet their specific research interests.

3 Predictive Impact

Once a feed definition is used to retrieve the relevant publications, the next challenge is to rank the results so that the user is presented with the most relevant papers first, rather than simply in chronologically order. Within the Meta Knowledge Graph, Article-Level Eigenfactor® (ALEF) values are calculated for each paper [5]. However, calculating non-zero ALEF scores on papers too recent to have been cited is not possible. To address this, Meta has developed a machine learning model which can be used to infer Eigenfactor® score for each recent publication. An *inferred* Eigenfactor® score is overwritten by the calculated score once enough citations are accrued. This inferred Eigenfactor® score is used as the sort value within the Meta feed.

4 Results Ranking

Many academic search engines allow for sorting of results on some sort of bibliometric impact, such as citation count or Eigenfactor®. However, this approach bias to older publications that have time to accrue higher citation counts, to the detriment of more recent publications. To address this within Meta, the query-independent inferred Eigenfactor® score is combined as a static rank value with query-dependent TF-IDF values [6] to calculate an overall score. Future work is planned to test combining these approaches with publication recency.

References

1. Meta, https://www.meta.org, last accessed 2019/06/26.
2. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J., BioBERT: pre-trained biomedical language representation model for biomedical text mining. arXiv:1901.08746 preprint (2019).
3. Chen, Q., Peng, Y., Lu, Z., BioSentVec: creating sentence embeddings for biomedical texts. The 7th IEEE International Conference on Healthcare Informatics (2019).
4. Campello, R.J.G.B., Moulavi, D., Sander, J., Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg (2013).
5. Wesley-Smith, I., Bergstrom, C.T., West, J.D.: Static Ranking of Scholarly Papers using Article-Level Eigenfactor (ALEF). arXiv:1606.08534v1 preprint (2016).
6. Robertson, S., Understanding inverse document frequency: on theoretical arguments for IDF", J Documentation 60(5), (2004).