Huma Zafar
University of Ottawa, Ottawa, Ontario, Canada

EXPLORING A MACHINE-GENERATED CONCEPT HIERARCHY THROUGH THE LENS OF 'NAIVE' CLASSIFICATION (Paper)

Abstract or Résumé:

Scholarly communication research often relies on comprehensive subject classifications to evaluate research produced within or across disciplines. Such use of classification systems is less related to information retrieval and more aligned with the types of knowledge discovery tasks described by Beghtol (2003) in her discussion of naive classification. In this thesis project in progress, we investigate the machine learning processes used to generate the Microsoft Academic Graph and OpenAlex subject classification systems to better understand how this classification supports knowledge discovery in a research evaluation context, and in what ways it might be made more effective in that context.

Introduction

While information retrieval (IR) is a frequently studied topic in the context of classification systems, the process of knowledge discovery – that is, the analysis of information resources to uncover *new* knowledge – can also benefit from well designed classifications (Beghtol, 2003). Such classification systems aim to organize resources in ways that primarily assist with analyzing those resources, rather than with finding and selecting resources (Beghtol, 2003). As an example of this usage of classification, Beghtol (2003) describes a research project where applying a specialized classification of religions based on their conceptualizations of the sacred helped surface certain similarities amongst groups of religions, leading to new theories about how and why monasticism arises in some religions and not in others (Dazey, 1994, as cited in Beghtol, 2003).

Bibliometrics and related meta approaches can be seen as another form of knowledge discovery, in which researchers seek new insights into the production of knowledge itself. Such research often relies on subject classifications of scholarly outputs to trace patterns within and across disciplines as well as to establish field-specific benchmarks. While traditionally, these knowledge organization systems were based on intellectual top-down approaches, increasingly, these subject classifications are derived and applied via algorithmic and machine learning processes (Golub, 2021). We believe that the growing misalignment between the design of these machine learning processes and traditional approaches to classification system design, particularly classifications targeted towards IR, has made them difficult to evaluate in depth theoretically.

In this work in progress, we investigate the generation and application of the subject classification deployed in OpenAlex, called OpenAlex concepts (formerly Fields of Science in Microsoft Academic Graph), from the perspective of knowledge discovery. The focus on knowledge discovery may yield new understanding of this classification's strengths and weaknesses with respect to bibliometrics and research evaluation. In turn, this may better help identify the effects of the machine learning components used in the derivation of OpenAlex concepts, and where, how, and with what justifications we can make improvements to the designs of such automated systems in general.

Literature Review

Classification for knowledge discovery has not been a widely discussed topic in the information studies literature since the concept was first discussed in detail by Beghtol (2003). Beghtol (2003) introduced the terms *professional* and *naive classification* to distinguish, respectively, between classification systems for the purpose of information retrieval, designed by information professionals and consciously guided by theoretical principles; versus classification systems "to help advance disciplinary knowledge" (Beghtol, 2003, p. 65), designed by domain experts. According to Beghtol, developing a naive classification to help study a set of materials involves first understanding the particular research questions that the classification should help answer, analyzing the materials to be classified in light of these questions, and then choosing the classes and subclasses to include in the classification system in order to meet the identified research needs (Beghtol, 2003). Though naive classifications may contain the same structural elements as professional ones (such as hierarchies and facets), these structures may be deployed in different ways when used in environments focused on knowledge discovery. (Beghtol, 2003)

A high-level discussion of the design of the MAG machine learning model is given by Shen et al. (2018), who describe the mechanisms for concept discovery, concept tagging, and concept hierarchy construction used by MAG. Since OpenAlex's concept hierarchy is taken directly from MAG's, this paper provides useful insight into the origins of OpenAlex's classification system as well, such as in the description of the algorithm used for constructing a concept hierarchy (Shen et al., 2018). OpenAlex (n.d.) further provide an overview of the OpenAlex concept tagger, as well as some of the technical design decisions behind its architecture and training process. However, they do not elaborate on the motivations for these decisions beyond helping their model perform similarly to the MAG concept tagger.

There have been many bibliometric studies conducted using MAG and/or OpenAlex data, demonstrating the frequent application of this classification system towards knowledge discovery tasks in the field of research evaluation. For recent examples, see Zafar et al. (2023), Huang et al. (2022), Liu et al. (2022), and Xu et al. (2024).

Preliminary Results

We have so far undertaken an initial investigation of the MAG concept hierarchy in order to compare the details of its construction with the process of developing naive classification systems, as per Beghtol (2003). When creating a naive classification within the scope of a single discipline and/or a single research project, one must first understand the research questions it is meant to help answer (Beghtol, 2003). In the case of a large-scale, general purpose system like MAG or OpenAlex, however, it is impossible to know *a priori* what investigations any researchers who use it may wish to conduct, and how they may want to use subject classifications; in this case, specific research goals cannot be used as a guiding principle for the classification system, and researchers may need to apply their own manipulations to the classification after it has been constructed. An example of such reformulation can be found in Huang et al. (2022), who removed seven concepts from the top-level of the MAG concept hierarchy after determining that these concepts contained a low degree of disciplinarity, and only then proceeded with their transdisciplinary analysis.

The second stage of generating a naive classification is to analyze the materials of the domain in order to determine a set of classes and subclasses for the classification system (Beghtol, 2003). In this case, the material to be classified exists in the domain of scholarly works, but the classes used in MAG come from the adjacent, but non-intersecting, domain of Wikipedia (specifically, Wikipedia article titles) (Shen et al., 2018). The concept of literary warrant mentioned by Beghtol (2003) appears to be applied in MAG's classification through the assumption that concepts which have their own Wikipedia article are significant concepts in general, but whether entries in a generalist encyclopaedia like Wikipedia accurately represent the scholarly literature covered in MAG and OpenAlex is debatable, and may also change over time.

Finally, structuring the MAG classification system into a hierarchy takes place only after the classes have been assigned to works (Shen et al., 2018). The construction of the MAG concept hierarchy uses the overlaps of concept assignments between papers in order to infer (pseudo-)hierarchical relationships between classes (Shen et al., 2018). This construction process thus has a reversed dependency between application of the classification system and the system's own construction; the classification's complete structure *follows from*, rather than dictates, its use.

The above illustrates how classificatory structures might be deployed differently in IR versus in knowledge discovery environments. Since the method of constructing the MAG hierarchy guarantees neither semantic coherence, nor transitivity of relationships (Shen et al., 2018), it does not abide by the usual (IR-centric) requirement that hierarchical relationships should represent either generic, instance, or whole-part relationships (National Information Standards Organization [NISO], 2010). However, it could be possible that the flexibility and sprawl of such a hierarchy might lend itself to discovering patterns across different fields of research better than a neater hierarchy would support.

In this talk, we plan to present an assessment of OpenAlex concepts through the theoretical lens of naive classification, and discuss the effect of its structure on knowledge discovery, particularly in the light of bibliometric approaches using OpenAlex.

References

Beghtol, C. (2003). Classification for information retrieval and classification for knowledge discovery: Relationships between “professional” and “naïve” classifications. *Knowledge Organization, 30*(2). 64-73.

Golub, K. (2021). Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, *59*(8), 702–719. https://doi.org/10.1080/01639374.2021.2012311

Huang, Y., Lu, W., Liu, J., Cheng, Q., & Bu, Y. (2022). Towards transdisciplinary impact of scientific publications: A longitudinal, comprehensive, and large-scale analysis on Microsoft Academic Graph. *Information Processing & Management, 59*(2). 10.1016/j.ipm.2021.102859

Liu, J., Chen, H., Liu, Z., Bu, Y., & Gu, W. (2022). Non-linearity between referencing behavior and citation impact: A large-scale, discipline-level analysis. *Journal of Informetrics, 16*(3). 10.1016/j.joi.2022.101318

National Information Standards Organization. (2010). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* (ANSI/NISO Z39.19-2005 (R2010)). ANSI/NISO. https://www.niso.org/

OpenAlex. (n.d.). OpenAlex: End-to-end process for concept tagging. Google Docs. https://docs.google.com/document/d/1q3jBlEexskCZaSafFDMEEY3naTeyd7GS/edit?usp=sharing&ouid=112616748913247881031&rtpof=true&sd=true

Shen, Z., Ma, H., & Wang, K. (2018). A Web-scale system for scientific knowledge exploration. *Proceedings of ACL 2018, System Demonstrations*, 87-92. 10.18653/v1/P18-4015

Xu, H., Liu, M., Bu, Y., Sun, S., Zhang, Y., Zhang, C., Acuna, D., Gray, S., Meyer, E., & Ding, Y. (2024). The impact of heterogeneous shared leadership in scientific teams. *Information Processing & Management*, *61*(1). 10.1016/j.ipm.2023.103542

Zafar, L., Masood, N, & Ayaz, S. (2023). Impact of field of study (FoS) on authors' citation trend. *Scientometrics*, *128*(4), 2557-2576. 10.1007/s11192-023-04660-2