The future of arXiv and knowledge discovery in open science

Steinn Sigurdsson
arXiv, Cornell Tech
ss3783@cornell.edu
& Pennsylvania State University
University Park, PA 16803

Abstract

arχiv, the preprint server for the physical and mathematical sciences, is in its third decade of operation. As the flow of new, open access research increases inexorably, the challenges to keep up with and discover research content also become greater. I will discuss the status and future of arχiv, and possibilities and plans to make more effective use of the research database to enhance ongoing research efforts.

1 Introduction

arχiv as of the time of writing contains $1,777,731$ e–prints across 158 categories in 8 different subject areas. The e–prints are distributed under license, and archived, free to the author and free to the reader. Distribution is fast, by daily e–mail blasts by category, RSS feeds, and direct web access. arχiv gets approximately a quarter of a million hits per hour, and currently receives about 15,000 new submissions per month. Each submission, upon acceptance, is assigned a unique arXiv:###.#####v# ID which is stable and maintains version control for revisions.

The average category has about five new primary submissions per day, but the larger categories have many dozens of new e–prints per day, and keeping up with the literature is accurately likened to trying to drink from a firehose. Finding the research you want and need seems to become harder as the tools to access the literature improve. Curating the flow of information and enabling discovery are critical tasks and expediting knowledge discovery will likely enable more rapid progress across a wide range of fields.

A peculiarity of the arχiv category system is that the submissions, each day, are ordered by time of submission, with the breakpoint being 2 pm eastern time each day. The e–prints submitted first after the 2 pm switch show up first, both in emails to subscribers and in web page listings. This leads to a well known effect that the e–prints first in this rank order are disproportionately cited compared to later ranked e–prints (Haque and Ginsparg, 2009, 2010). While this provides a form of knowledge discovery, and one that undoubtedly correlates with some attributes of the researchers who wrote and submitted the e–print, it is not an optimal technique for efficient discovery.

Additional discovery is enabled by the sorting of e–prints into *categories* within their *subject* areas. While this is primarily done by authors at the time of submission, the ultimate choice of categorization is done by arχiv moderators, with the assistance of automated tools. arχiv reserves the right to set the primary category in which the submissions appear, and to assign or remove secondary categories or later cross listings.

Moderation of Categories: The moderation process within arχiv currently occurs in three primary stages. On submission the e–print is processed and checked for a number of technical issues. Approximately 10% of e–prints have some error and are referred back to the author (not counting minor errors which are generally fixed during submission). The median time from a submitting author logging into the system and a submission being complete is 34 minutes.

During submission an automatic classification system scans the paper and recommends a choice of categories. For most submissions the choice of primary category matches that of the author, in other cases the system may recommend a different category. The author may choose to accept the new recommended category, or add it as a secondary category in which to list the submission, or they may proceed with their original category and ignore the system recommendation.

Proceedings of the First Workshop on Scholarly Document Processing, pages 7–9

https://doi.org/10.18653/v1/P17

Each category has an assigned moderator, with 194 voluntary moderators covering the range of subject categories. Some categories have two or more moderators, generally the high volume categories. At any given time there are gaps in moderation, sometimes covered on an *ad hoc* basis by volunteer super–moderators, who may have responsibility for multiple categories. The moderators are generally senior PhD researchers who are active in the field they are moderating. The ideal moderator is someone who was checking the distribution in "their" category every morning anyway, someone who wants to see first what is new.

The moderators have primary responsibility for the choice of categories, and whether a submission is released, held for further checks, or rejected. A small fraction of submissions get held for an extended period (currently about 0.5% are on hold for two or more weeks). Extended holds mostly occur due to coordination issues between moderators, including moderators who are out of action, disagreements on choice of categories, or policy questions. Submissions may be rejected for being out of scope, not being the type of content arχiv accepts, or not meeting the threshold for standard of acceptance in that subject and category. arχiv makes substantial effort to maintain consistent standards across categories and subject areas, but there are some differences in approach and style across academic fields, and where these cross is often where papers are put on hold or rejected. arχiv is a curated collection, it is not a general repository. arχiv is not the internet. In order to remain useful to its community of users arχiv curates content for relevance and interest, while trying to avoid gatekeeping and active refereeing of content. This means there are always borderline cases. The borderline cases are generally not important in the aggregate, but they are important to the individual authors, and there is a strong motivation not to exclude original or innovative approaches through overly strong filtering. However, there is always some border between accepting and rejecting, and wherever that border is, there are always edge cases which end up being judgement calls. Moving the border does not resolve the issue, it merely moves which submissions are borderline.

2 Classification

Papers submitted to arχiv are run through a natural language processing classifier (Ginsparg et al., 2010). Currently three classifiers are operational: the original full text classifier from Ginsparg; a beta version of a more general broad classification package, including full text, run asynchronously; and a new fast metadata classifier developed by Papers with Code. Note that arχiv submissions come with very sparse metadata. Demanding large amounts of metadata provided by the author puts a burden on the author during submission and discourages use of arχiv for rapid distribution of research.

Moderators see the classifier score, and a paper with a high score in a particular category not selected by the author may be queued up for consideration by the moderators of that category, for selection either as a primary category or as a secondary choice of category. A few percent of submissions typically get some category changes, often the addition of one or more secondary categories for listing.

Category changes also lead to significant fraction of holds of submissions, both when a moderator deems a submission unsuitable for the choice of category, or when moderators disagree amongst themselves about the choice of category. Orphaned submissions, those rejected from all choices of primary categories, may be rejected as out of scope. Moderators will often recommend alternative possible choices of primary, or recommend secondary categories for submission.

The classifiers are imperfect, in particular when trying to determine a fit for the smaller categories, even after balancing, and some broad categories are treated as exception cases. Training the classifier is an iterative process, and more work is needed.

3 Knowledge Discovery

Ultimately researchers and other readers want to discover the latest research that is relevant to their interests, and to find other relevant results, novel methods, complementary insights or other useful or interesting knowledge. A lot of discovery comes from finding your lane and staying in it, the categories provide useful silos for a significant fraction of researchers and push most of the directly relevant science to that community. Beyond the silo, searches of the literature are useful for discovery, but are often constrained by what can be indexed for searching and how the search algorithm keys in on search terms. Improved search algorithms and federated cross platform searches generally improve prospects for discovery.

arχiv currently has formal relations with several

entities to expedite cross platform discovery, including INSPIRE, ADS, DBLP, Semantic Scholar and Google Scholar. We are working to improve discovery including author disambiguation. Author ID services such as ORCID and Institutional Identifiers like ROR also expedite searches. An ambition of arχiv is to provide custom delivery of new submissions beyond the current categories, to include among other options, author selection, types of content, and inclusion and exclusion by keywords and relevance.

Balancing this impetus is the danger of loss of discovery by browsing, the serendipitous discovery that came when browsing a physical journal and finding a surprise article adjacent to the one you were seeking, or a topical book you were not aware of shelved next to those you browsed. Refined and narrow searches limit surprises. Sometimes what you are looking for is not the "known unknowns" but rather the "unknown unknowns". It is tempting to consider providing a small fraction of random or semi-random search results in searches in the hope of triggering the rare discovery of an unknown.

A more formal process may be more efficient and likelier to succeed, and arχiv is interested in pursuing knowledge discovery techniques, including knowledge graphs and novel techniques for finding relevant results that are not adjacent to the research area being searched. There are very large benefits to finding an existing solution to an experimental problem, a new computational technique making your modeling tractable, a statistical or mathematical method making your problem solvable, or the novel theoretical insight from a different subfield.

More broadly we want to find emerging new directions of research, even before those doing the research realize there is an emergent effort which is headed in a new direction, to see disparate subfields converge into new synergistic research opportunities, and adjacent subfields diverge to nucleate new areas of research. These are hard problems, but exciting and with very high potential for discovery and speeding up research.

arχiv core functionality is to get the paper to the reader, but quantity has a quality all of its own (Clement et al., 2019). Bulk downloads of content for natural language processing, machine learning and other aggregate exploration has been enabled for some time through Amazon's S3, with the user paying. arχiv has now partnered with kaggle to provide bulk access to arχiv contents, providing both aggregate metadata, and access to processing full text. The kaggle dataset is updated periodically and is free to use. Text retrieval of any particular paper still goes through arxiv.org.

4 arχiv Labs

arχiv has set up a framework for us to work on a range of issues with external partners through arχiv Labs, https://labs.arxiv.org. Currently arχiv Labs includes the arχiv Bibliographic Explorer, a new collaboration with Papers with Code to link papers and code (https://paperswithcode.com), and the CORE Recommender (Knoth et al., 2017) (https://core.ac.uk).

arχiv Labs is committed to open source, and partners working through the framework are expected to abide by the general arχiv principles. We are interested in working with individuals or groups on third party services, as well as more structured services that could be brought in–house and run from the arχiv side as services to our users, or even part of our core operations. We are in discussion about several other projects.

References

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. *Computing Research Repository*, arXiv:1905.00075.

Paul Ginsparg, Paul Houle, Thorsten Joachims, and Jae-Hoon Sul. 2010. Last but not least: Additional positional effects on citation and readership in arxiv. *Computing Research Repository*, arXiv:1010.2757.

Asif-ul Haque and Paul Ginsparg. 2009. Positional effects on citation and readership in arxiv. *Computing Research Repository*, arXiv:0907.4740.

Asif-ul Haque and Paul Ginsparg. 2010. Last but not least: Additional positional effects on citation and readership in arxiv. *Computing Research Repository*, arXiv:1010.2757.

Petr Knoth, Lucas Anastasiou, Aristotelis Charalampous, Mattero Cancellieri, Samuel Pearce, Nancy Pontika, and Vaclav Bayer. 2017. Towards effective research recommender systems for repositories. *Computing Research Repository*, arXiv:1705.00578.