Table of contents

1 New Gen Research Indexing

Elevator pitch: Our research corpora is the collection of human insight. Certain aspects were thoughtfully constructed (i.e. citations and indexing algorithms), others were products of circumstance, and some are relics of the past. What does the ideal scientific literature, corpora, and search engine interface look like (in the generative age)? With the main goal being that researchers should be able to find the work that is the most valuable to them. An auxiliary goal being that automated reasoning systems (ML systems) can best take advantage of the literature to speed up the discovery process. In other words, what would a neo-Google Scholar look like?

1.0.1 Audience

The (key) audience in mind is the Google Scholar team; a secondary audience would be the Elicit AI team.

2 Overview

Core Q: How is the discovery, evaluation, and use of research changing in the generative era? How can we redesign research indexing to meet these new challenges and opportunities? The ethos here is that finding non-indexed things is much more difficult (*what would it look like and under what circumstances?)

Novel contributions: In the rigorous, quantitative estimate of an implemented framework, detailed sketch of what an implementation would look like, the review of the literature because it's been unexplored. Solid, concrete directions for work (practical, useful details and caveats).

- We could enter a new age of search indexing (where it's dynamic, concept-based, and decentralized): imagine that you had influence scores for training data inputs.
- Synthetic or low-value research is much easier to produce now and indexers should work on promoting high-value, novel content to those who are using a search engine.

- Tagging papers by type (survey, experimental, empirical, etc.).
- Privacy-preserving activity logging and knowing where sources are found from.
- Memorized completions aren't as useful here.
- Centralized platforms own critical metadata (e.g. citation graphs) that make decentralized or open alternatives difficult to access.
- Keeping the chat history for certain projects centralized and then citing them.
- A new age of search for research papers that relate to the concepts in the papers themselves: how do you unlock semantic search?
- Integrating OpenReview and peer review into the search process as a means of including confidence metrics into the works themselves.
- Questions: What is Google Scholar's dominance and over what time period?
- Having citations for Datasets (NeurIPS announced their Datasets and Benchmark track in 2022 and we should incentivize the development of datasets and part of that comes with attribution and ownership)

2.1 Desiderata/Ideas

- To be able to see how knowledge builds on each other by seeing how different concepts and artifacts build on each other and integrate into one another.
- Transparency on decisions related to an artifact (removals, etc.)
- Linking preliminary workshop papers to main journal and conference proceedings. The author being able to group different research projects.
- Seeing when components of a paper are re-used/cited, such as graphs and charts (This would require persistent identifiers for these graphs.) The tools that people use for making graphs and what is currently possible.
- Having datasets linked on Google Scholar as well (dynamic (API-retrieved data) compared to static (CSV, instantiated datasets))
- Citing software and libraries (*release versions as metadata)
- Centralization (linking to OpenReview and the comments that a document would receive)
- Research corpora demography: We're now able to see the derivative cited works from a paper but how would this work for quotations (key points of an article that are recited as proxies for different concepts and their popularity)
  - This would allow for detailed similarity checks and being able to assess how original a work is compared to the broader literature
- Having hashes for files on the arXiv as digital fingerprints
- Encouraging meta-analyses and visualizations of the google scholar dataset (see the ISBN visualization bounty)
- In 2014, there were 160 million documents indexed [re: Wired article]. *Get the current number and compare across repositories (i.e. Elsevier, ArXiv, etc.)

2.2 GScholar History & Background

- Developed by Alex Verstak and Anurag Acharya (both still at Google)
- Going through their most recent work and GScholar history:
  - Verstak and Acharya filed a patent in 2017, approved in 2019 (?), (on behalf of Google) on a system for identifying the primary source of a document when there are mulitple versions of a document. As it is, internet search engines don't have supervisory control which means that a multi-authored paper should be submitted by one of the editors to ensure that there is only one version and the history can be tracked as such. Their system involves an authority conferral for each version of a document obtained from different sources (i.e. different publishers, pre-print servers, repositories, etc.). The authoritative source is given based on the priority order of the source (with an algorithm consisting of parameters like page rank, citation, keywords, publisher exclusivity rights, and so on) and length qualification. The scores will indicate a ranking of priority in addition to meeting a length qualification. *Re SJ: It would be interesting to reconfigure the parameter weights and have this be visible to the user or allow them to reorder search results based on this.
  - Seeking approval from publishers to allow Google to crawl their repositories and thereby allowing them to circumvent paywalls. While JSTOR (the largest archive of research then) provided scans of articles it didn't "read the articles." Initially, they could acquire the first page of the article and get author information and the abstract (I'm unsure if this is still the case today) *In what ways do current machine-readings come short?
  - Alerts were an area of focus about a decade ago. What would be the meaningful things to be alerted on? See the derivative projects that came from a paper? *As the user, being able to refine the alerts that you receive. ## Auxiliary
- Some issues like if you web scrapers that surpass paywalls then the scientists don't get paid (but that never really mattered: could run an experiment and see if that's possible – jailbreaking)

2.3 Remarks and Next Steps

- Wikidata relies on the Blazegraph which isn't being mantained

3 Bibliography

- OpenReview Paper
- The Gradient: Text Embedding Inversion

- arXiv: Copyrights and Licenses for Academic Works
- arXiv: Foundational Research Paper
- Wired article on GScholar History, 2014