IBM Use-case, Experiences, and Position Statement for AI-CONTROL

Authors: Heiko Ludwig, Nirmit Desai

IBM Research

hludwig@us.ibm.com, nirmit.desai@us.ibm.com

IBM is engaged in advancing the state-of-the-art in building generative AI capabilities in the open, anchored on the IBM Granite[1] family of open-weight models as well as recipes used in preparing the training datasets in Data-prep-kit[2], both available under Apache 2.0 license. A key part of this mission is to self-govern the acquisition, processing, and usage of public data sources that are crawled. The usage spans the use cases of pre-training, fine-tuning, instruction-tuning, and RAG.

Challenges IBM experiences in responsibly crawling public data

1. Web servers enforcing unpublished rate limits; while large Web sites throttle crawlers in a benign way, rejecting requests by time out or return codes, smaller sites occasionally get overwhelmed by performant crawlers and experience service disruption. This neither serves the crawler nor the Web site operator.
2. Lack of standardized mechanism to register and authenticate crawlers and track their reputation; some Web sites require crawler naming but might not be familiar with a user agent.
3. Lack of a standard taxonomy of AI-based use cases and purposes for web crawled data; in discussion with data owners we find that some are amenable to make data available for pre-training but not for RAG, for example, or vice versa, or at different conditions, requiring document-level attribution for RAG but not for pre-training.
4. Lack of standard specification of terms of use and licenses associated with web content that is consumable by crawlers. This is more of a problem in jurisdictions outside the US that don't have a fair use doctrine or a nonstandard way of opting out, for example, of the TDM exception in the EU.
5. Lack of objective auditability of which crawler requested and downloaded what web content, for what purpose, and when. A standard log format helps documenting under which conditions a document was acquired.

[1] https://huggingface.co/ibm-granite
[2] https://github.com/IBM/data-prep-kit

6. Lack of incentives for the Web publishers to freely allow crawling of their content. At present, Websites are concerned about not being fairly compensated. Ideally, they would be incentivized to provide data, e.g., by pointing to a royalty service.

Proposed changes, with robots.txt as a starting point

A declarative specification in a well-known location in the way robots.txt is mostly used is a suitable mechanism and we propose it as a baseline of a specification discussion. The TDM Reservation Protocol[3] as proposed by W3 provides alternate approach conveying opt-out and condition information in a http response header but we estimate that this will lead to much network overhead, crawlers asking for content that is rejected. It appears to be more efficient if a site-level specification defines what is permissible or excluded.

Robots.txt does not specify a crawling purpose. Either a robots.txt extension or a complimentary specification following a similar approach could extend this. While we might imagine many use cases for data – and more to come – we also want to keep it simple. One approach would be to define the following:

- ai-training: all training-related activity, including pretraining, fine/instruction/etc. tuning, and other processes producing a model
- ai-knowledge-base: the use of content as a knowledge base for an AI system, e.g., RAG.
- ai-synthetic: the creation of a derivative data product such as synthetic data based on the content of a Web site.
- ai-other: all other AI use cases of the content.

Each purpose can be complemented by a condition of use as defined in a pointer to a license. Ideally, this would be a standard license such as the Creative Commons family of licenses with clear definitions of the expectations of the data user. Bespoke licenses are not helpful for automated interpretation, besides interpretation by an LLM. This is mostly relevant in jurisdictions that do not have a clear legal doctrine of use.

Crawl delay specifications are not always a good mechanism and practical to implement by distributed crawlers. Alternatively, a specification such as requests/minute would be easier to implement using a counter by crawler and Web operator and potentially provide a better frame of reference for a specifying Web host operator, who often manage hosting capacity by the request velocity they experience or support.

Complementing the specification, a simple, domain-level audit format would help Web site operators request, when their domain was crawled. It could comprise start and end time, and the number of files retrieved. This could be made available on specific request as a trust-building measure. The robots.txt-type file could specify an email address to which this summary could be sent.

[3] https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/#abstract