The Missing Science of Knowledge Curation

(Improving incentives for large-scale knowledge curation)

Praveen Paritosh
Google, pkp@google.com

Abstract

Dictionaries, encyclopedias, knowledge graphs, annotated corpora, library classification systems and world maps are all examples of human-curated knowledge resources that have been highly valuable to science as well as amortized across multiple large-scale systems in practice. Many of these were started and built even before a crowdsourcing research community existed. While the last decade has seen unprecedented growth in research and practice in building crowdsourcing systems to do increasingly complex tasks at scale, many of these resources are still woefully incomplete—lacking coverage in languages and subject matter domains. Moreover, many knowledge resources needed to fill other semantic gaps for artificial intelligence systems simply don't exist or aren't being built. Why? I argue that we don't have the right incentives, and that in order to improve the incentives, we have some fundamental scientific questions to answer. While building a large knowledge resource, we have little more than intuitions when it comes to estimating the reusability, maintainability, and long-term value of the effort. These make it difficult to fund or manage such projects, often requiring herculean personalities or fortunate businesses. Building or expanding a resource is often not seen as "sexy," which results in lack of resources to answer those questions in any principled manner. These problems begin to outline a new science of curation, making progress on which could help improve the discussion around and funding for building sorely needed knowledge resources.

ACM Reference format:
Praveen Paritosh. 2018. The Missing Science of Knowledge Curation (Improving incentives for large-scale knowledge curation). In Companion of The Web Conference 2018, Lyon, France, April 2018 (WWW'18), 2 pages. DOI: https://doi.org/10.1145/3184558.3191551

The unreasonable effectiveness of human curation

Large-scale, human-curated knowledge resources such as dictionaries, encyclopedias, knowledge graphs, annotated corpora, world maps, and library classification systems are substantial investments of resources. The cost of building and maintaining these resources has been amortized across multiple large-scale systems, where they have been shown to be disproportionately more valuable than their cost. For example, linguistic knowledge resources such as treebanks and inventories such as wordnet are used by most information and question answering systems. Factual knowledge resources such as knowledge graphs and structured world maps have been key for question answering systems such as Watson, search engines, and for entity and relation extraction body of research. Despite these successes, there are some widely held myths about human curation.

Myths shaping the incentives for human curation

I believe that we have failed to create proper incentives to foster investments in building and scaling these knowledge resources. Here are the narratives that have shaped these broken incentives, which should be revisited given the recent successes in large-scale curation and crowdsourcing:

- Expanding a knowledge resource is not new science.
- Machine-learned resources are more scalable.
- Long-term value of knowledge resources is impossible to quantify.

Let us examine each one of these myths more closely.

"Expanding a knowledge resource is not new science"

Enumerating by hand is considered a weak, if not the weakest, scientific theory. However, expanding it in a reusable and extensible way is! Just as no cartographer starts from scratch, it is the scientific obligation of a knowledge resource to be easy for future users and developers to extend it, as well as to be interoperable with other resources. Can others reuse and maintain it? Does it share identifiers and vocabulary with other resources? What other best practices can help make the content of knowledge resources more enduring?

The Pareto principle might be good news to those who care about the content, suggesting that a small catalog can capture a large number of cases in the world. But current incentives focus disproportionate research on the long tail of edge cases, while never cataloging the head with its high explanatory power per unit cost. There is a lot of science in modeling and in building elegant formalisms to house the data that will forever remain empty if we don't value collecting the data.

WWW 2018, April 23-27, 2018, Lyon, France.

ACM ISBN 978-1-4503-5639-8/18/04. https://doi.org/10.1145/3184558.3191551

"Machine-learned resources are more scalable"

Just to take a case study, the research community (both academic and industrial) has spent far more resources in automatically expanding resources, such as Freebase, than in building them in the first place. There are early results in expanding the easy parts, using the human-curated core as a bootstrap. Even after substantial investment in automatically curating and expanding Freebase, most successful knowledge resources are almost entirely human curated. In addition, machine-learned resources critically depend on human-curated resources such as Wikipedia.

"Long-term value of knowledge resources is impossible to quantify"

One-off solutions always fare better in this calculus of immediate utility-based funding than something that will be more expensive up front but amortize later. This is a very tricky one, as currently seen, most knowledge resources are in service of some systems. Thus their evaluation and success is intrinsically tied to the system's success. This makes the evaluation both myopic to a single utility versus amortization across new forms of usage that come in the wake of the existence of the resource. The ways in which dictionaries and wordnets are being put to by IR/QA systems could not be easily imagined when it was being built.

However, we have learned to counteract this in fields, such as software development, that benefit from amortization. We have learned to fund and incentivize infrastructure building and capacity planning of assets such as factories and data centers. To the extent that there are more uncertainties in characterizing the specs and risks of knowledge resource projects than other long-term investments, we need best practices and supporting science for estimating current and future value of knowledge resources.

We need a new science of knowledge curation

The issue with incentives is knowing when things are "better." Understanding and scaling curation is about characterizing the content, not the form. Having strong empirical or theoretical understanding of these issues would help reduce the risk in making larger investments.

While there is a rigorous foundation and understanding of formalisms used to write knowledge—whether it be triples, frames, logics, or controlled vocabularies—we don't know much about the best practices of how to write a billion assertions in a reusable and maintainable manner in any of those formalisms. Currently, most of this knowledge of practice is tucked away in curation guidelines and the minds of the curators.

Unlike the foundations of the formalisms, which can be done in the artificial universe of mathematics, knowledge curation is a human task and thus any understanding of this requires venturing into fields as diverse as survey design, linguistics, psychology, sociology, the so-called soft sciences. In other words, there are hard scientific problems about characterizing the curation and the content of large-scale knowledge resource building efforts.

In order to improve the incentives, we need a science of knowledge curation that addresses these misconceptions and answers questions about how to measure and compare reliability, coverage, explanatory power, progress, utility, and the long-term value of knowledge resource building efforts.

References

[1] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 28th ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD 2008), Vancouver, Canada.

[2] Chang, N., Paritosh, P., Huynh, D., & Baker, C. (2015, June). Scaling Semantic Frame Annotation. In LAW@ NAACL-HLT (pp. 1-10).

[3] Ferrucci, David, et al. "Building Watson: An overview of the DeepQA project." AI magazine 31.3 (2010): 59-79.

[4] Han, S., Dai, P., Paritosh, P., and Huynh, D. (2016). Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control. ACM Transactions on Intelligent Systems and Technology (TIST), 7(4), 56.

[5] Ipeirotis, P., and Paritosh, P. (2011). Managing Crowdsourced Human Computation. Tutorial presented at WWW 2011. In Proceedings of the 20th International World Wide Web conference, Hyderabad, India.

[6] Josephy, T., Lease, M., Paritosh, P., (2014). Crowdsourcing at Scale workshop report. AI Magazine, vol 35, No. 1.

[7] Kochhar, S., Mazzocchi, S., and Paritosh, P. (2010). The Anatomy of a Large-Scale Human Computation Engine, In Proceedings of Human Computation Workshop at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2010, Washington D.C.

[8] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data, ACL, Association for Computational Linguistics.

[9] Paritosh, P. (2012). Human Computation Must Be Reproducible. In Proceedings of CrowdSearch: Crowdsourcing Web Search at the 21st International World Wide Web Conference (WWW), Lyon, France.

[10] Riezler, S. (2014). On the problem of theoretical terms in empirical computational linguistics. Computational Linguistics, 40(1), 235-245.

[11] Sameki, M., Barua, A., and Paritosh, P. (2016). Rigorously Collecting Commonsense Judgments for Complex Question-Answer Content. In Third AAAI Conference on Human Computation and Crowdsourcing.

[12] Welty, C. A., & Jenkins, J. (1999). Formal ontology for subject. Data & Knowledge Engineering, 31(2), 155-181.

[13] Zhang, A. X., Culbertson, B., and Paritosh, P. (2017). Characterizing Online Discussion Using Coarse Discourse Sequences, In 11th AAAI International Conference on Web and Social Media (ICWSM).