The Knowledge-Remixing Bottleneck

Groth, P.T.

IEEE Intelligent Systems, 28(5), 44-48

Published version: http://dx.doi.org/10.1109/MIS.2013.138

Link VU-DARE: http://hdl.handle.net/1871/50554

The Knowledge-Remixing Bottleneck

Paul Groth
Network Institute, VU University Amsterdam

Traditionally, the knowledge-acquisition bottleneck has been a core problem in intelligent systems. How do we get information into an intelligent system so that it can reason and operate over it?

Over the past five to ten years, we've seen how the Web has been central to attacking this problem: for example, the use of Web corpora has enabled large-scale natural language processing,[1] the emergence of community-derived knowledge bases such as DBpedia and Wikidata (see www.wikidata.org), and the application of Web-based data in order to play Jeopardy.[2] However, to construct these knowledge bases, we purposely mix, munge, and clean the data. Indeed, in one study, analysts spent 60 percent of their time in data preparation.[3] This remixing process is central to data science and is a key aspect of performing Web Science. In doing so, though, we remove nuance, context, and provenance. As Danah Boyd has pointed out, "working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth."[4]

In trying to understand the Web and the feedback loop of social and technical constructs that creates it, it's vital for us to be able to interrogate how the knowledge we make our assertions upon was created. This includes not just the computational methods used but also the *decisions* that underlie the use of those methods. However, transparency is only one part of the story; the ability to "pull apart" and reassemble our knowledge bases, revisiting and substituting the decisions that led to their creation, would allow us to reuse and repurpose them for new kinds of analytics. Here, I attempt to concretely formulate this problem and look at potential research directions for addressing it.

Decisions Are Central

One of the primary examples of Web Science is the study of political movements and how they're mediated by the Web—in particular, through social media websites such as Twitter. The canonical example is the use of Twitter within the Iranian Elections in 2009—the so-called "Twitter Revolution." Of the subset of papers that looked at Twitter usage during this event, Devin Gaffney's work is of particular interest,[5] as it argues for a Web Science methodology based on automated large-scale analysis as compared to prior studies, which adopt an anthropological methodology—choosing a curated subset of websites and analyzing these (almost) manually. To perform these automated analyses, a knowledge base is constructed from a corpus of tweets. For instance, a retweet network is constructed by extracting when a tweet was retweeted, either by querying the Twitter API or by doing text extraction using the convention of "RT" within a given tweet. Gaffney, like many other authors, interprets this retweet network as representing influence—the more people retweet your tweets and the more those tweets cascade through the network, the more influential you are. The influence of a given person can then be measured using handy network statistics such as centrality. For example, Gaffney found that "persiankiwi, mousavi1388, tedchris" were all influential on Twitter during the 2009 Iranian election using such statistics.[5]

However, this result hinges, critically, upon the decision to interpret the retweet network as influence. Indeed, when we construct knowledge bases from Web data, such decisions are necessary to structure and make use of the data. But, we need to recognize that *other decisions are possible*. Indeed, a retweet can have many meanings.[6] For example, one could interpret a retweet network as primarily a mechanism of communication—where the hubs are essential broadcast points. The point is not which interpretation is better; instead, it's that the decision is often lost, not made explicit, or the justifications for it aren't given.[7]

This lack of transparency around the decisions made during knowledge base creation becomes even more pernicious as we move farther away from a single data source where a particular decision can be connected more readily back to its origin.

Hidden Decisions: Integration Exacerbates the Problem

When we construct a knowledge base from more than one dataset, we not only decide which part of the data we want to use and how to interpret it, but also how the integration algorithm performs. Because of the complexity of the preparation process, these decisions can easily become hidden both within the extraction and integration procedure as well as the data itself. An example is in the construction of large knowledge bases using Wikipedia, such as DBpedia.

Extraction and Integration Procedures

DBpedia relies on 10 different extractors for acquiring knowledge from Wikpedia. Each extractor makes different assumptions about the underlying data itself.[8] For example, the mapping-based infobox extractor leverages a hand-built ontology based on the 350 most commonly used English-language infobox templates. The decisions about the arrangement of that ontology aren't reported in the paper[8] that describes the extraction process for DBpedia.

Furthermore, DBpedia uses multiple external classification schemes to help make the extracted data queriable. Each of these classifications is based on its own construction techniques and processes. One of these, YAGO,[9] is automatically constructed from the combination of Wikipedia infoboxes and another database, Wordnet.[10] In YAGO's extraction routine for constructing its class hierarchy, the authors use Wordnet synsets as upper classes. They then map Wikipedia classes (extracted from the Wikipedia category hierarchy) as subclasses of some Wordnet synset. This mapping relies on the frequency of word occurrences within a synset provided by Wordnet. The decisions to use this heuristic are described in the paper about YAGO, [9] but tracing this information back from DBPedia requires some effort. Also, it isn't apparent from the YAGO paper how word frequency is calculated within Wordnet. None of these decisions are accessible when querying DBPedia, and even when reported in various papers or online, it requires some effort to track it back.

I should point out that this isn't a criticism of these projects—indeed DBPedia, YAGO, and Wordnet are extremely transparent—but it's a clear indication that the state of the art in our work practice is far from ideal, and that many decisions remain hidden.

Underlying Data

Decisions aren't just hidden within integration and extraction procedures, they're lost in the data itself. Wikipedia clearly reflects the notions of the community that contributes to it.[11] This community, as with most crowd-sourced sites, is dominated by a set of self-selected contributors. The top 10 editors by number of edits contributed 86 percent of the valuable content.[12] Furthermore, most contributors don't contribute equally across the site, instead focusing on a small number of interest areas.[13] Likewise, the data within Wordnet reflects the views of the trained psycholinguists who built it. Finally, Wikipedia is also influenced by the tools that are used for its construction (that is, bots).[14] All these contributors have their own unique biases and points of view, which impact their decisions about what to include and not include in Wikipedia.

Thus, DBpedia is the consequence of layer upon layer of decisions and interpretation of underlying data sources. It combines self-selected community data, expert-produced linguistic information, and other integrated knowledge bases all using hundreds of decisions made in its integration procedure. Tracing back all the decisions and their interconnections that led to DBpedia is surely a difficult proposition.

Is This a Problem?

As a resource for analysis, DBpedia and other Web-sourced knowledge bases are extremely useful. For example, these resources are being integrated into text analytics pipelines (such as http://spaziodati.eu), which is a key technique for Web science, data science, and computational social science. Such incorporation has ramifications when analyzing a corpus of tweets using such text analytics; the resulting analysis would contain not just the interpretations and decisions of the analysis, but also all of the aforementioned decision points. This might hide particular biases or errors. For instance, in a political situation, a jurisdiction might not be recognized because of the decision of a community member not to include it in their classification. In other domains, such ramifications are also evident. In the biological sciences, the decision to map a protein entry in one database to a gene entry in another could improve recall but might lead to incorrect results.[15]

This lack of transparency can, to some degree, be solved through careful and painstaking forensic work. From a systems perspective, the inability to *revisit* the decisions that were made in knowledge base construction and *reuse* portions of the creation pipeline is a more serious problem. Indeed, this lack of ability to revisit different decision strategies slows the process of analytics. For instance, if someone wants to reuse a retweet network but instead interpret it differently, does a person need to recreate that network or can it be used as is?

The Knowledge-Remixing Bottleneck

With these issues in mind, the knowledge-remixing bottleneck can be defined in two parts:

- the difficulty in tracking the decisions by which a knowledge base was constructed from multiple data sources; and
- the inability to repurpose parts of a knowledge base construction procedure.

This is termed a *bottleneck* because these activities are feasible, but they're far from being automated at any scale. The word *remix* is chosen intentionally, because it connotes not just reuse or repurposing of knowledge, but the ability to change decisions about how knowledge is integrated. In some sense, we can make an analogy to the way editors work with music or video—they can revisit their decision to cut, splice, select, and fade different sources together in a certain way. This ability to examine and adaptively modify their decisions is critical for an editor's work practice.[16] Currently, Web and data scientists are missing this capability. To revisit and remix knowledge bases requires extensive manual effort, ranging from reading papers to understanding someone else's code—and we should work to eliminate this manual effort.

Addressing the Bottleneck

A first step to address this bottleneck is to look at our own work practices as Web and data scientists. We can be more faithful in documenting our design decisions and linking them to our analysis and integration code. Furthermore, we can publish the code we use and do our best to make it reusable. This approach has been advocated elsewhere[17,18] and is surely an important part of Web science practices. However, documentation isn't enough; instead, we should focus on the tools we create to build and remix knowledge bases.

One area to look at is the use of explicit descriptions of our remixing pipelines in terms of computational workflows. Indeed, a common paradigm for construction and editing of video is termed *node-based compositing*, which arranges the process of editing in a workflow from inputs to outputs where each particular edit can be inspected and modified at runtime, thus changing the resulting video composition. David De Roure and Carole Goble have argued that in fact method, as expressed in a workflow, should be the central artifact in Web science research.[19] However, our current tools (scripts) don't yet have the affordances to quickly build analysis and integration pipelines. We do see some signs of the effectiveness of workflow in the growth of frameworks for data analysis, such as Hadoop and Signal/Collect, but we're still a long way off from Nuke for knowledge base creation (Nuke is a node-based editor for video compositing; see www.thefoundry.co.uk/products/nuke).

Workflows can be complemented by the automated collection of data provenance from the computing environment and exposed using standards such as World Wide Web Consortium's Provenance specification (W3C PROV).[20] This interoperability of data provenance is critical for being able to track back across systems. The ability to span systems begins to tackle the problem of aggregate knowledge bases being constructed from other aggregate knowledge bases (such as DBpedia). However, this approach is premised on the adoption of the standard. Luckily, examples are beginning to emerge. DBpedia has begun to assert which Wikipedia page its data was extracted from using PROV. The Git2PROV service (http://git2prov.org) allows any Git repository to be exposed as PROV-formatted provenance. A great example of provenance enabling the inspection of decisions is the Karma Data Integration tool (see http://isi.edu/integration/karma), which lets users investigate what automatic data integration decisions the tool made. This capability was used to hand curate links from the Smithsonian American Museum to Wikipedia.[21]

Both workflows and provenance can help to better understand how knowledge bases are constructed and repurpose parts of the construction process. However, both focus on more algorithmic or computational decisions. Interpretation decisions are left to the side. Better systems for capturing the decision making of humans as it relates to knowledge base construction are needed. For example, in the construction of the Never-Ending Language Learning (NELL) knowledge base,[22] human feedback is used to help guide decision making about what facts to include. The question is can we capture those decisions systematically and link them to the construction process? The encoding of such decisions (the "why") is essential for understanding the context behind a knowledge base. Wikidata has begun to address this by enabling specific references to each of the facts included in its knowledge base. Here, an interesting task would be to connect a particular interpretative decision to the set of computational processes that implement it. This would allow a person to revisit these interpretive decisions and potentially replace them with a new set of computational processes that reflect a different interpretive decision. Capturing these sorts of decisions is challenging when those involved are non-experts or layman.

A final area of interest is to combine the aforementioned transparency with opinion mining on the sources themselves.[23] This would attempt to draw out the implicit biases with the underlying data and make them explicit.[24] Obviously, this process would have to be documented again using approaches like the Knowledge Diversity Ontology (see http://kdo.render-project.eu)**.**

Web and data science often rely on knowledge bases constructed from the Web. By leveraging the sheer scale of the Web, we're beginning to solve the problem of knowledge acquisition. However, the construction of these knowledge bases leads to a new problem—the opacity of the construction process itself. It's difficult to determine the decisions made during the data preparation and integration process due to a lack of transparency. Moreover, the assumptions and interpretations embedded within a given dataset are rarely available in a structured fashion. This lack of transparency hinders repurposing knowledge bases and makes it more difficult to leverage the full capability of these Web-based knowledge acquisition systems. As we continue to perform Web science studies, it's imperative that we're clear about what knowledge those studies are based on and the decisions manifested in them. Beyond this awareness, as a community, we should look at developing intelligent tooling that helps us pull apart and remix our knowledge bases.

References

1. A. Kilgarriff and G. Grefenstette, "Introduction to the Special Issue on the Web as Corpus," *J. Computational Linguistics*, vol. 29, no. 3, 2003, pp. 333–347.

2. J. Chu-Carroll et al., "Textual Resource Acquisition and Engineering," *IBM J. Research and Development*, vol. 56, nos. 3–4, 2012, pp. 4:1–4:11.

3. NASA, A.40 Computational Modeling Algorithms and Cyberinfrastructure, tech. report, NASA, 19 Dec. 2011.

4. D. Boyd and K. Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Comm., & Society*, vol. 15, no. 5, 2012, pp. 662–679.

5. D. Gaffney, "#iranElection: Quantifying Online Activism," *Proc. Web Science 2010: Extending the Frontiers of Society On-Line*, 2010. **http://journal.webscience.org/295/**

6. D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," *Proc. Hawaii Int'l Conf. System Sciences*, IEEE, 2010; doi:10.1109/HICSS.2010.412.

7. D. Freelon, "On the Interpretation of Digital Trace Data in Communication and Social Computing Research," *J. Broadcasting & Electronic Media*, to be published, 2013; http://dfreelon.org/wp-content/uploads/2008/06/dfreelon_tracedata_preprint_JOBEM.pdf. **//Do you have an update on this? Has it been published? I'm just wondering if we can direct the reader to the volume, issue, and page numbers.//**

8. C. Bizer et al., “DBpedia—A Crystallization Point for the Web of Data,” *J. Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, 2009, pp. 154–165.

9. F.M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A Large Ontology from Wikipedia and WordNet,” *J. Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 3, 2008, pp. 203–217.

10. G.A. Miller, “WordNet: A Lexical Database for English,” *Comm. ACM*, vol. 38, no. 11, 1995, pp. 39–41.

11. F. Flöck, D. Vrandecic, and E. Simperl, “Towards a Diversity-Minded Wikipedia,” *Proc. ACM 3rd Int’l Conf. Web Science*, ACM, 2011; www.websci11.org/fileadmin/websci/papers/112_paper.pdf.

12. R. Priedhorsky et al., “Creating, Destroying, and Restoring Value in Wikipedia,” *Proc. 2007 Int’l ACM Conf. Supporting Group Work*, ACM, 2007, pp. 259–268; http://doi.acm.org/10.1145/1316624.1316663.

13. R. Almeida, B. Mozafari, and J. Cho, “On the Evolution of Wikipedia,” *Proc. Int’l Conf. Weblogs and Social Media*, 2007; www.icwsm.org/papers/paper2.html.

14. S. Niederer and J. van Dijck, “Wisdom of the Crowd or Technicity of Content? Wikipedia as Socio-Technical System,” *New Media & Society*, vol. 12, no. 8, 2010, pp. 1368–1387.

15. C.Y.A. Brenninkmeijer et al., “Including Co-Referent URIs in a SPARQL Query,” *Proc. 4th Int’l Workshop on Consuming Linked Data*. **CEUR-WS.org, 2013, http://ceur-ws.org/Vol-1034/BrenninkmeijerEtAl_COLD2013.pdf**

16. E. Laurier, I. Strebel, and B. Brown, “Video Analysis: Lessons from Professional Video Editing Practice,” *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 9, no. 3, 2008, article no. 37 .

17. R.D. Peng, “Reproducible Research in Computational Science,” *Science*, vol. 334, no. 6060, 2011; pp. 1226–1227.

18. C. Goble, “The Reality of Reproducibility of Computational Science,” IEEE eScience Conf. keynote presentation; www.sysmo-db.org/node/64.

19. D. De Roure and C. Goble, “Anchors in Shifting Sand: The Primacy of Method in the Web of Data,” *Proc. Web Science 2010: Extending the Frontiers of Society On-Line*. **http://journal.webscience.org/325/**

20. P. Groth and L. Moreau, eds., PROV-Overview: An Overview of the PROV Family of Documents, W3C Working Group Note NOTE-prov-overview-20130430, World Wide Web Consortium (W3C), Apr. 2013; www.w3.org/TR/prov-overview.

21. P. Szekely et al., “Connecting the Smithsonian American Art Museum to the Linked Data Cloud,” *Proc. 10th Extended Semantic Web Conf.*, LNCS 7882, Springer-Verlag, 2013, pp. 593–607.

22. A. Carlson et al., “Toward an Architecture for Never-Ending Language Learning,” *Proc. Conf. Artificial Intelligence*, AAAI, 2010; www.cs.cmu.edu/~acarlson/papers/carlson-aaai10.pdf.

23. B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval*, vol. 2., nos. 1–2, 2008, pp. 1–135.

24. F. Giunchiglia, V. Maltese, and B. Dutta, “Domains and Context: First Steps Towards Managing Diversity in Knowledge,” *J. Web Semantics: Science, Services and Agents on the World Wide Web*, vols. 12–13, 2012, pp. 53–63.

***Paul Groth** is an assistant professor in the Department of Computer Science at the VU University Amsterdam. This article was supported by the Data2Semantics project in the Dutch national program COMMIT. Contact him at p.t.groth@vu.nl or on Twitter @pgroth.*