What if the primary goal of the web was to foster curiosity?

Praveen Paritosh

Google, San Francisco, USA

Abstract. People go to the web to satisfy their curiosity. The web contains resources that can help: articles, videos, tutorials, online communities, and online courses, among others. In analogy to the semantic web proposal, which was motivated by a desire to structure the web to be more understandable and usable by machines [Berners-Lee, Hendler, and Lassila, 2001], we raise the question: How would we rethink the web with the primary goal of fostering and satisfying human curiosity? We propose the *curiosity web*, based on the intuition that the meaning of resources, such as articles, books, and videos, can be expressed in terms of the questions they address [Paritosh and Marcus, 2016]. It has three representational elements: *curiosity*, a semantic primitive for an abstracted question or information need with a URI and textual content in multiple languages; *relationships between curiosities*, such as relevant or prerequisite; *relationships between curiosities and resources*, such as addresses or satisfies. The goal of the curiosity web is to provide an exoskeleton for organizing information by the curiosities they address. The curiosity web is a dual of existing semantic networks and knowledge graphs [Collins and Quillian, 1972; Sowa, 2006; Hillis, 2004]. Instead of focusing on describing meaning using analytic primitives and compositions of knowledge, this approach represents meaning of resources in a holistic manner, through the curiosities it addresses.

Keywords: Knowledge Representation, Semantic Web, Knowledge Graph, Wikipedia

1 What are some challenges for existing models of organizing knowledge?

In areas of information of wide public value – such as healthcare, culture, lifestyle, and arts – social, experiential, and subjective knowledge abounds. Consider the information needs of an expectant mother over the course of the pregnancy, to childbirth, to raising and parenting the child. While we have made great strides in representing and organizing structured knowledge about people, places, and things, not as much progress has been made in organizing such social knowledge.

For example, Wikipedia, knowledge graphs, web markup, and wikidata, to name a few, can tell us a lot about the movie Blade Runner: when it was released, who produced it, the cast, the screenwriter, the ratings on IMDb and other sites,

and so forth. These are verifiable facts that we have consensus on. But there are more complex curiosities about the movie, such as the following:

- Why was the movie named Blade Runner?
- What makes it a cult classic?
- What are the differences between the various different cuts of the movie?
- What does the origami represent in the movie?

These are legitimate curiosities. The state of the art in knowledge graphs and structured data does not have the schemata to represent the knowledge required to understand, let alone answer, such questions. In addition, we lack scalable methods of curating such social knowledge, as it involves subjectivity and multiple, even conflicting, perspectives. Aggregating and ranking such knowledge goes beyond the existing operationalizations of consensus and authority. Sociologists have called this the *social stock of knowledge* [Berger and Luckman, 1991], and library scientists have named this *everyday life information seeking* [Spink and Cole, 2001], but our communities have not paid enough attention to these domains of knowledge.

There is content addressing this valuable class of information needs, but it is hard to find. It is prevalent in unstructured text and media on forums and blogs and reviews and email groups and social media discussions and comments and videos and books. This knowledge is trickier to represent, curate, and evaluate. It is full of subjectivity, ambiguity, disagreement, and differing perspectives. Nevertheless, this social stock of knowledge is vital to addressing the information needs of patients coping, parents distraught, children curious, and solving real-world problems.

2 Why is social knowledge hard to organize?

Knowledge bases such as Wikipedia have too strict of an epistemological position (including objectivity, notability, and verifiability) that makes it impossible to talk about most of the social knowledge that we are interested in – this knowledge literally has no home! And so it stays in the halfway houses of QA websites and forums. While Wikipedia is curated and organized purposefully for a wide readership, much of social knowledge on the web is found knowledge created as a by-product of human interaction in online communities. This content is not finessed or curated for consumption by someone who is not part of this community and discussion. More structured knowledge graphs [Bollacker et al., 2008] lack the primitives such as verbs, adjectives, and adverbs, as well as the notion of context and perspective [Guha, 1991], which makes it impossible to represent social knowledge.

Knowledge graphs, topic maps, concept maps provide exoskeletons for organizing the map of knowledge. What are they missing? Topics are too coarse and are not helpful for describing the aspects that people care about, and the connectivity between topics does not include any kind of conceptual or curricular organization. And lastly, in our communities of computer scientists and

engineers, we have an epistemological inclination toward precision, truth, and accuracy, which is limiting for characterizing social knowledge.

3 How do we organize knowledge for and by curiosities?

The curiosity web is an approach to organizing knowledge by curiosities. The underlying intuition is that we can express what a document does, for the purposes of information consumption, by connecting it to the questions it addresses. We choose to start with questions, since they are shorter and simpler than answers. In addition, we can disagree about the answers more than we can about the questions, providing a more stable foundation for the knowledge in answers. Curating the curiosity web involves curating URIs for curiosities and the relationships between curiosities (such as related or prerequisite), and relationships between curiosities and resources (such as addresses or satisfies).

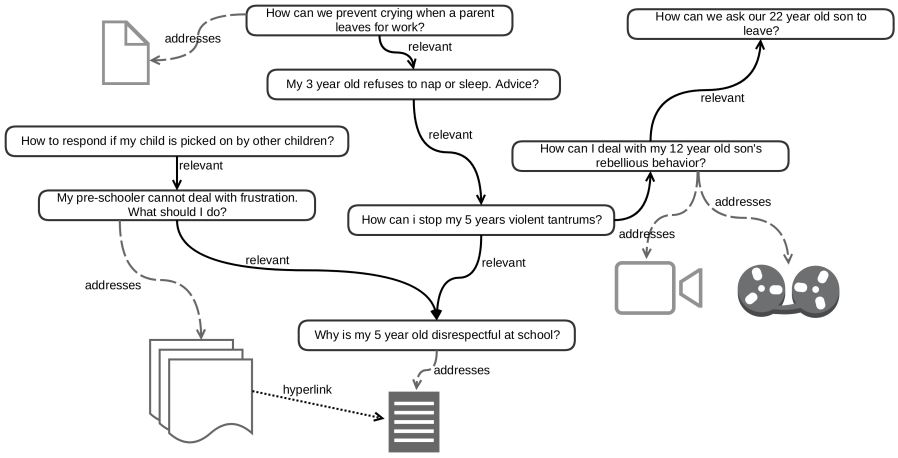

Fig. 1. A small slice of the curiosity web focusing on the questions of parents and resources related to them

First, we propose a ***curiosity*** as a first-class semantic primitive for representing information needs. A curiosity has a stable URI, and textual descriptions and elaborations in multiple languages. Just the Wikipedia curation process for the topic for the book Alice in Wonderland abstracts away across specific editions and copies of the books, curiosities in the curiosity web will be curated by abstracting from questions. Thus, a curiosity might further point to one or more questions on the web that ***surface*** that curiosity.

Second, the curiosity web contains relationships between curiosities. The most simple relationship between two curiosities might be that they are ***relevant*** to each other, that is, someone interested in one might find the other

useful. In addition, there might be curricular relations such as prerequisite or enables, and other relationships such as temporal order, granularity, or level of detail.

Third, the curiosity web contains relationships between curiosities and resources. The simplest relationship is that a resource ***addresses*** a curiosity. We can imagine additional relationships such as evokes, satisfies, and so forth. Below we show a hypothetical example of a small part of the curiosity web around curiosities related to parenting.

4 How will we build and maintain the curiosity web?

There are existing examples of organizing knowledge using questions, for example, the widespread usage of frequently asked questions (FAQs) to organize community knowledge on forums and Usenet [Hammond et al., 1995]. Another starting point for questions on the web might be in schema.org/Question markup [Guha, Brickley, and Macbeth, 2016]. Further, community QA sites including stackexchange.com merge similar questions into one thread so as to not fragment content across multiple threads. Another example is amazon.com, which has a crowdsourced product FAQ page attached to every product on its website. Questions are a compact human-understandable and curatable knowledge representation for information needs. Building and maintaining the curiosity web is feasible; here are some models addressing parts of the task:

1. *Community driven:* Imagine Curiositypedia, built on top of Wikipedia community principles to curate curiosities and connections between them.
2. *Webmaster driven:* Imagine schema.org markup for articles and content by webmasters to make their content more available to user questions and search engines.
3. *Machine driven:* Using technologies such as statistical machine translation, annotate resources on the web with curiosities they address.
4. *User driven:* As users consume content, they can help annotate it with their questions that the content addressed, much like collaborative tagging efforts [Golder and Huberman, 2006]. Everybody can be a curator, and the curiosity web gets better as users leave trails!

There might be benefit in trying different methods as they might produce vastly different results being helpful for different use cases. For example, Wikipedia, Knowledge Graphs, Linked Open Data, RDF, schema.org, open information relation extraction approaches, they all inform the analogous semantic web proposal.

5 How will the curiosity web help foster human curiosity?

Curiosity is the urge to know. Curiosity is fascination with the unknown. A curiosity is a desire to know or learn something; for example, one might want to know how to make an omelette fluffier, or why a five-year old cant shake off

temper tantrums, or the meaning of it all. This might be motivated by a need to solve a problem or explore new domains. This might surface in forms such as questions posed to friends or queries to search engines. It surfaces as the question mark in a sentence, the queries to search engines and social media, the questions for elders and librarians, the communities of inquiry, and at a larger scale, the collective processes of accumulating knowledge, such as science and spirituality.

The obvious unknown is the known unknown – the location of the restaurant your friend recommended, the latest on the California fires, etc. These are unknowns just on the surface of what we know and need. Resolving them quickly improves our lives and makes us expand our surface of known. Then there is the unknown unknown, things that are so beyond our stock of the known that we wouldnt even know what to ask. And yet, our curiosity is fascinated with these unknowns – the desire to travel, to seek new experiences, to go outside our comfort zone – there are some real risks to such a fascination with this distant unknown. Even though we dont have a specific question (usually) when we visit a new place, that journey and our interactions there could teach us new things, or help us see what we knew in a new light, and lead us to new questions.

The curiosity web can make it easier for the creators, curators, and consumers of content to annotate content with curiosities they address. The hypothesis is that questions are a natural, systematic, and human-understandable representation for organizing knowledge. Others who later have similar curiosities do not have to figure out how their curiosities map to some taxonomy, or know the right terms to pose queries, but can more directly rediscover such content. In addition, being able to connect curiosities together allows us to build and follow the trails of knowledge that the Memex machine dreamed of [Bush, 1945].

References

1. Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250).
3. Bush, V. (1945). As we may think. The atlantic monthly, 176(1), 101-108.
4. Collins, A. M., and Quillian, M. R. (1972). Experiments on semantic memory and language comprehension.
5. Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema. org: evolution of structured data on the web. Communications of the ACM, 59(2), 44-51.
6. Hammond, K., Burke, R., Martin, C., and Lytinen, S. (1995, February). FAQ finder: a case-based approach to knowledge navigation. In Artificial Intelligence for Applications, 1995. Proceedings., 11th Conference on (pp. 80-86).
7. Hillis, W. D. (2004). Aristotle (the knowledge web). Edge Foundation, Inc, 138.
8. Paritosh, P., and Marcus, G. (2016). Toward a comprehension challenge, using crowdsourcing as a tool. AI Magazine, 37(1), 23-30.
9. Sowa, J. F. (2006). Semantic networks. Encyclopedia of Cognitive Science.