Touring the World Wide Web

Jim Hendler, Heng Ji and Mei Si
Rensselaer Polytechnic Institute

The dominant way to find content on the Web today is via a search engine -- but it wasn't always like that. In the earliest days of the Web, users who downloaded the early browser-prototypes would go to some starting place on the Web, such as Cern's TheProject.html[i] and start exploring. Much of the joy of the Web was finding new and different things one didn't know existed. With Websites being few and far between, the joy was often in the serendipitous discovery of a new area or technique that one didn't know existed. To provide a metaphor for this interaction, in an article in 1992 Jean Armour Polly introduced the term "surfing" to refer to the act of interacting with the Web. In explaining the choice of this term she wrote:

> *"I weighed many possible metaphors. I wanted something that expressed the fun I had using the Internet, as well as hit on the skill, and yes, endurance necessary to use it well. I also needed something that would evoke a sense of randomness, chaos, and even danger."* (Polly, 1992)

In short, the journey on the Web was about discovering what was out there, not on finding specific content.

In an effort to reintroduce the idea of exploring the web, rather than searching, we are developing a new technique that combines state of the art techniques in Information Extraction (Huang and Ji, 2015), Semantic Web data integration (Hendler, 2015), Cognitive Computing (Hendler and Ellis, 2014), and Interactive Storytelling (Si and Marsella, 2014). Here's how:

- Using the new "living information extraction" technique, we will be able to create a "never-ending extractor" which will be pulling from web documents information about entities and events, and the relationships between them. The new system can work in a dynamic node, and does not need human annotated samples for training, but it works best if there are a number of known relationships between pages to build off of.
- The Semantic Web provides a number of known relationships between pages on the Web in a number of domains. Using general knowledge sources, like dbpedia and Yago, and specialized knowledge sources, like the data from musicbrainz, the reviews from Yelp (which have semantic annotations) and even the Open Graph of Facebook (which is available in a semantic web format), provides a jumpstart for the language extraction. However, the Semantic Web relates pages, but doesn't have any sort of "understanding" of what is on the pages.
- Cognitive Computing, in this case using the architectural ideas that underlie Watson, can allow us to have a better way of accessing information about the entities found on the Web and finding other information about the same entities using various kinds of search and language heuristics. The system can have various kinds of "frames" about entities and try to fill them – what are the interesting things to know about a person, a musical event, a product, and many

other domains is an extension to the kind of question-answering that the original Jeopardy-playing Watson performed. This allows us to have more organized information, rapidly generated, about the entities being explored. However, given a large graph of entities (even the organized linked-open data cloud has information about billions of things), how do we choose what to display next? If the best we can do is provide links, all of the above isn't much better than choosing a page and clicking from there.

- Interactive storytelling techniques are being explored to take information in the kind of "knowledge graph" resulting from the above, and tailoring the presentation to a user using storytelling techniques. It is aimed at presenting the information as an interesting and meaningful story by taking into consideration a combination of factors ranging from topic consistency and novelty, to learned user interests and even a user's emotional reactions. The system can essentially determine "where to go next" and what to do there in the organized information as processed above..

This ambitious project will try to unify these different technologies being developed by computer scientists, cognitive scientists, web scientists, ethnographers, and artists, and use them to provide what is, at first, an interesting (and we hope amusing) experience. Over time, studying how humans use such a system, what different kinds of user categories might find interesting, how this might differ across cultures, and how such a system might be used as the core of many other innovations is another project goal.

In this talk, we will briefly overview the different technologies, show why we believe this really could work, and demonstrate some of the first examples of doing this for various kinds of information sources on the web.

References

Hendler, J., Data Integration for Heterogeneous Datasets, *Big Data, 2(4),* December, 2014 *(*doi:10.1089/big.2014.0068)

Hendler, J. and Ellis, S., Why Watson Won, 2015 - http://www.slideshare.net/jahendler/why-watson-won-a-cognitive-perspective

Huang, L. and Heng Ji et al.. 2015. Living Information Extraction. Submission to Proc. the 2015 Conference on Empirical Methods on Natural Language Processing.

Polly, J. A. (1992) "Surfing the Internet 1.0" in *Wilson Library Bulletin* June 1992.

Si, M., & Marsella, S. C. (2014). Encode Theory of Mind in Character Design for Pedagogical Interactive Narrative, *Advances in Human-Computer Interaction*, Volume 2014 (2014), Article ID 386928,

[i] A copy of the 1992 version of this page can be found at http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html