

Social Meaning on the Web: From Wittgenstein to Search Engines

Harry Halpin and Henry S. Thompson, *University of Edinburgh*

One could hypothesize that the biggest question for the Web is whether multiple agents in a decentralized information space can share meaning via the use of uniform resource identifiers (URIs), such as *http://www.example.org*. On the hypertext Web, this bet was trivial; most of the time a URI

would identify a Web page by virtue of allowing access to the Web page itself. However, even in the Web's earliest stages, URIs were for more than just accessing Web pages: they united the previous disparate protocols of the Internet into a single seamless and smooth space of information, where any network-accessible object could be given a URI. Yet, as Tim Berners-Lee noted in his keynote speech to the World Wide Web Conference in 1994, "To a computer, then, the Web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them."¹ The goal of the Semantic Web, then, is to give URIs to "real objects and imaginary concepts" as well as to the "relationships between them." However, there is a fly in the ointment: a Web browser cannot simply access a real object like the Eiffel Tower via HTTP. So, the original question of what a URI identifies, which we could answer earlier by trivially accessing a

Web page, transforms into the open question of how agents can determine what non-Web-accessible thing a URI in the Semantic Web identifies in a decentralized manner. This is the defining problem for the evolution of the Web into the Semantic Web.

Theories of Reference and Semantics

There are two opposing, yet plausible, stories about how these URIs in the Semantic Web get their meaning. In the first story, as advocated by Berners-Lee and others with a background in Web architecture, a URI gets its meaning from its owner. This seems to be a plausible enough story, since in the original hypertext Web, this is precisely how URIs worked: the owner of the URI had the authority to host Web pages or other network-accessible objects on the host whose name began the URI itself. However, the Semantic Web presents a disturbing question: what, if anything, should be accessible from these URIs for real-world things

Relevance feedback from hypertext Web searches considerably improves the performance of a Semantic Web search engine, creating a mutually beneficial relationship between the hypertext and Semantic Webs.

and imaginary concepts? Obviously, a straightforward response would be to host some sort of accurate description that is accessible (perhaps via redirection) from the URI, such as a picture of the Eiffel Tower and some data about the Eiffel Tower in a language like RDF (Resource Description Framework, the primary knowledge representation language of the Semantic Web). The success of linked data illustrates that the hosting of these descriptions over HTTP is critical. However, there are edge cases. What if the description is ambiguous? What if there are multiple descriptions for what appears to be the same thing?

The second story, as advocated by Pat Hayes, the primary author of the formal semantics of RDF, is precisely the inverse: It is the descriptions—in particular, interpretations of the model given by the formal semantics of those descriptions—that define the possible referents of any URI. So, it doesn't matter what the owner of the URI *thinks* his URI identifies. The descriptions, not the owner, are the determining factor, and the best the owner can do is give access to descriptions that communicate his referential intention. Therefore, Hayes considers it essential to ditch the vague word “identify,” as used in URIs, and distinguish between the abilities of URIs to access and to refer. Although access is constrained by Web architecture, Hayes argues that reference is absolutely unconstrained except by formal semantics, so “the relationship between access and reference is essentially arbitrary.”² Furthermore, informal descriptions are generally considered ambiguous and difficult for automated processing by machines. Therefore, formal semantics for knowledge representation languages are created that can precisely define possible inferences, and hence referents.

Disagreement on this point crystallized as a debate between Berners-Lee and Hayes, in which Berners-Lee put forward the hypothesis that a URI “identifies one thing,” and Hayes responded that URIs are always ambiguous, because the interpretation given by the formal semantics of the Semantic Web assigns the same URI to different individuals in different models.

This debate can be considered a return to a long-standing debate in the philosophy of language on the meaning of names, if we think of URIs as just names on the Web. Hayes' position can be considered analogous to

Can we define a theory of senses for the Web on the basis of use, rather than mere logical inference or the supposed intention of the URI's owner?

Bertrand Russell's *descriptivist theory of reference*, in which a name is considered functionally a set of descriptions that can refer to different individuals.³ As an alternative to the descriptivist theory of names, Saul Kripke proposed the *causal theory of reference*, wherein the referent of a proper name is given by an act of “baptism” and then causally transmitted through time, so that a name refers to a unique individual over all possible worlds.⁴ This position seems similar to that of Berners-Lee, with the creation of a URI, backed up by ownership of the relevant domain name, being analogous to the act of baptism. Yet, according to this story

about URIs, if a URI exists for the Eiffel Tower itself, it doesn't matter if a picture of blue cheese is accessible from that URI, or if accurate RDF descriptions are hosted there at all. The URI means what the owner says or thinks it does, and the descriptions are merely ancillary.

Social Semantics and Wittgenstein

With the rise of Web 2.0, it is clear that a new contender for semantics is on the scene: social semantics. To return to our analogy to debates in philosophy of language, there has long been a third position, wherein names are given their meaning by social and linguistic practice. This position was first articulated by Ludwig Wittgenstein in his “late” period, in a repudiation of his earlier strongly logicist viewpoint. On this account, the meaning of any expression, including a URI, is grounded not only in its formal truth value or referent but in its socially and linguistically constructed “sense.” The notion of sense can be reconstructed so that it can denote more than just truth values. In Wittgenstein's view, sense can be construed in terms of the socially grounded norms that are necessary to grasp the use of a name within a language by its users. Thus arises the infamous slogan, “meaning is use.”⁵

Although it is a topic of much controversy, the notion of sense is intuitive: when you look up a word in a dictionary, you get several different dictionary definitions of the word, which are often the different senses of the word's use. The question then becomes, can we define a theory of senses for the Web on the basis of use, rather than mere logical inference or the supposed intention of the URI's owner? The number of senses needs to be open ended, as the content of the Web constantly changes; it cannot

be simply a closed, finite number of senses in a standard dictionary.

Wittgenstein's most revolutionary concept was the notion of a "form of life," such that to "imagine a language is to imagine a form of life."⁵ His term *language-game* is "meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life."⁵ What would be the form of life of the Web? Submitting queries to search engines and browsing the results certainly plays a central part therein. For the Semantic Web to succeed, the meaning of a URI should have its formal meaning supplemented by its social meaning. From a purely pragmatic standpoint, given the historic difficulties of classical AI, it might make more sense for the Semantic Web to ally itself with the phenomenally successful practice of information retrieval in addition to knowledge representation.

The discipline of information retrieval descended directly from Wittgenstein himself via Margaret Masterman. One of the six students in Wittgenstein's course that became *The Blue Book*, Masterman was exposed to the fundamental concepts that were the foundation of Wittgenstein's *Philosophical Investigations*.⁶ Two decades later, in the late 1950s, Masterman founded the Cambridge Language Research Unit, where Karen Spärck Jones laid the foundations for information retrieval. Information retrieval, especially its data-driven and statistical methodology, is basically neo-Wittgensteinian philosophy of language given computational flesh. Search engines such as Google are at least implicitly neo-Wittgensteinian, as are certain techniques like tagging.

However, the core problem of information retrieval remains that

natural-language query terms are inherently ambiguous and usually result in too much data being retrieved, even from the Semantic Web.⁷ For example, if a user wants to find a URI to denote the Eiffel Tower and accordingly types the keywords "Eiffel Tower" into a Semantic Web search engine such as Falcons,⁸ they receive a plethora of URIs and Semantic Web documents in RDF that mention the Eiffel Tower. Which one should they use? How can we increase the likelihood that the "best" URI for the Eiffel Tower is at the very top of the list of search results?

The core problem of information retrieval remains that natural-language query terms are inherently ambiguous and usually result in too much data being retrieved.

Relevance Feedback between the Hypertext Web and the Semantic Web

By observing the behavior of users in selecting certain Web pages from the results of ordinary hypertext search (that is, not Semantic Web search), the precise information the user is interested in can often be detected via techniques from information retrieval and even natural-language processing. Then the hypertext Web pages can be used to approximate the additional social meaning of the query terms, and this in turn can be used in combination with machine-learning techniques to disambiguate Semantic

Web URIs while maintaining the socially necessary ambiguity. We see this as putting these Semantic Web URIs in what Ricardo Baeza-Yates has called a "virtuous cycle" with the hypertext Web.⁹ Although we lack the space to describe the algorithm in full (it is available in other works⁷), in essence, both the hypertext Web page and the Semantic Web data are converted into a "bag of words" and then compared using well-known techniques from information retrieval—in particular, *relevance feedback*. Relevance feedback takes some known relevant documents and uses them to "expand" the user's usually short one- or two-word query by adding possibly relevant words from the relevant documents.

In our experiment, detailed more thoroughly elsewhere,⁷ we chose 200 queries from a log of queries submitted to Microsoft's Live search engine, evenly divided between people and place names on the one hand, and names of "abstract concepts" (identified as such by WordNet) on the other. We defined relevancy as whether or not a Web page or Semantic Web document was about the same thing as the query. We determined this, in turn, by asking experimental subjects whether or not either the Web page or the Semantic Web document expressed accurate information about the thing queried. For each query, we used Yahoo Search to retrieve the top 10 Web pages, and Falcons to retrieve the top 10 Semantic Web documents. Using a forced binary-choice paradigm, we showed the subjects each of the 10 hypertext Web results and 10 Semantic Web results (formatted using Disco Hyperdata Browser) and had them choose whether or not the Web page or RDF was relevant to the initial query. Each hypertext Web page and Semantic Web document was judged by three subjects. For relevance judgments over both Semantic Web results and Web page results,

$\kappa = 0.5724$ ($p < 0.05$, 95 percent confidence interval [0.5678, 0.5771]), indicating the rejection of the null hypothesis and moderate agreement between subjects. How many queries had a relevant Semantic Web result as their top result? The Semantic Web results had 76 (58 percent) top-ranked relevant results. This means that almost half of the time, in response to a query to the Semantic Web, the top-most result would not be relevant.

To improve this, we used several vector-space models to implement relevance feedback, where the primary parameter to be varied was the window size of top-frequency, nonzero words to be used in the vectors for both the query model and the document models. We ran baselines with no relevance feedback. We give full details of the results of the most popular relevance-feedback algorithms and parameters in another publication.⁷ Given in terms of mean average precision in Figure 1, the results show that relevance feedback as implemented by the famous Inquiry information retrieval system¹⁰ with a window size

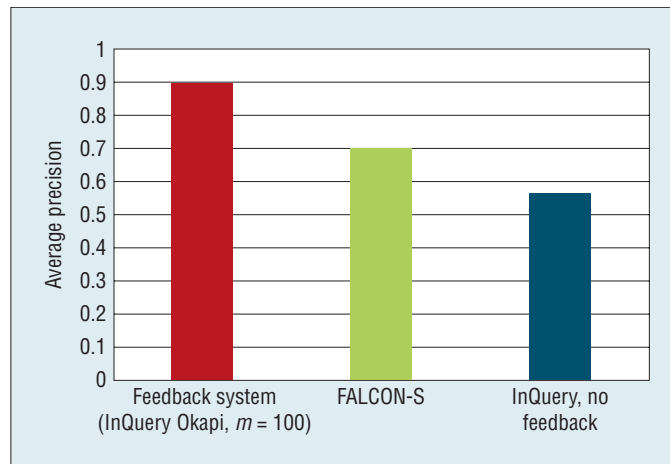


Figure 1. Summary of best average precision scores: relevance feedback from hypertext to Semantic Web. A system that uses feedback from the hypertext Web (red) outperforms both the deployed Semantic Web search engine (green) and the baseline information retrieval system without feedback (blue). The baseline information retrieval controls for all parameters except the presence of feedback from the hypertext Web, unlike the deployed Semantic Web search engine.

of 100 had the best performance. Using the Wilcoxon signed rank test, the relevance feedback system had a mean average precision of 0.8914 ($p < 0.05$), significantly better than both Falcons (0.6985, $p < 0.05$) and the same parameterized algorithm without relevance feedback (0.5595, $p < 0.05$). It appears that relevance feedback between relevant hypertext Web pages does help in reranking Semantic Web search results.

What does this average precision increase mean? In practice, it means that after reranking, 89 percent of the Semantic Web results now have a relevant URI in first place, a large 31

percent increase over the baseline. This level of increase in precision transforms Semantic Web searching from being fraught with errors to being good enough for daily use. For example, one application of this technique would be, for a given hypertext search query, to automatically retrieve relevant structured Semantic Web data, such as key facts and contact information, and display this data to the user with the hypertext results. The massive numbers of hypertext queries can automatically discover separate high-quality, structured information on the Semantic Web, allowing a mutually beneficial relationship between the hypertext and Semantic Web to emerge.

Relevance feedback works because the information in the relevant hypertext Web pages about something like the Eiffel Tower also forms an accurate model of information that should be in a Semantic Web description of the Eiffel Tower. There is only one Web, and hypertext Web pages and the Semantic Web are both about real-world objects and imaginary concepts that can be described using fragments of natural language, ranging in complexity from tagging to full sentences, to formal Semantic Web descriptions. All of these forms of description are mutually complementary because they are all grounded in the same forms of life, as Wittgenstein would put it, even if the form of life on the Web is the use of search engines with keywords.

If the Web is to be considered a first-rate subject for scientific inquiry, issues of social meaning cannot be thrown by the wayside as being somehow unscientific. Arguments

THE AUTHORS

Harry Halpin is a fellow of the World Wide Web Consortium (W3C), serving as chair of the Social Web Incubator Group and as staff contact for the RDB2RDF Working Group, which is working on a standard to export relational data to the Semantic Web. His research interests include the intersection of information retrieval and structured data as in the Semantic Web. He has a PhD in informatics from the University of Edinburgh.

Henry S. Thompson is a reader in the School of Informatics at the University of Edinburgh and a staff member of the World Wide Web Consortium (W3C), where he works in the XML Activity and serves on the W3C Technical Architecture Group. His research interests include the semantics of markup, XML pipelines, and more generally understanding and articulating the architectures of the Web. He has a PhD in linguistics from the University of California, Berkeley.

about names and meaning that have emerged in philosophy cannot and should not be avoided. As Berners-Lee put it, “We are not analyzing a world, we are building it. We are not experimental philosophers, we are philosophical engineers.”¹¹ Finding and giving meaning to URIs on the Semantic Web can be achieved by appeal to the social semantics implicitly manifested in the searching and browsing behavior of ordinary users of the hypertext Web. ■

References

1. T. Berners-Lee, “World Wide Web Future Directions,” plenary talk, Int’l World Wide Web Conf., 1994, www.w3.org/Talks/WWW94Tim.
2. P. Hayes and H. Halpin, “In Defense of Ambiguity,” *Int’l J. Semantic Web and Information Systems*, vol. 4, no. 2, 2008, pp. 1–18.
3. B. Russell, “On Denoting,” *Mind*, vol. 14, no. 4, 1905, pp. 479–493.
4. S. Kripke, *Naming and Necessity*, Harvard Univ. Press, 1972.
5. L. Wittgenstein, *Philosophical Investigations*, trans. G.E.M. Anscombe, Blackwell Publishers, 2001.
6. Y. Wilks, “A Personal Memoir: Margaret Masterman (1910–1986),” *Language Cohesion and Form*, M. Masterman, Cambridge Univ. Press, 2005.
7. H. Halpin and V. Lavrenko, “Relevance Feedback between Hypertext Search and Semantic Search,” *Proc. Semantic Search Workshop at the World Wide Web Conf.*, 2009, http://km.aifb.uni-karlsruhe.de/ws/semsearch09/semse2009_27.pdf.
8. G. Cheng, W. Ge, and Y. Qu, “Falcons: Searching and Browsing Entities on the Semantic Web,” *Proc. World Wide Web Conf. (WWW 08)*, 2008, <http://www2008.org/papers/pdf/p1101-cheng.pdf>.
9. R. Baeza-Yates, “From Capturing Semantics to Semantic Search: A Virtuous Cycle,” *Proc. 5th European Semantic Web Conf. (ESWC 08)*, LNCS 5021, Springer, 2008, pp. 1–2.
10. J. Allan et al., “INQUERY and TREC-9,” *Proc. 9th Text Retrieval Conf. (TREC-9)*, US Dept. of Commerce and NIST, 2000, pp. 551–562.
11. T. Berners-Lee, message to P. Hayes on W3C Technical Architecture Group mailing list, 16 Jul. 2003, <http://lists.w3.org/Archives/Public/www-tag/2003Jul/0158.html>.

cn Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

computing now
ACCESS | DISCOVER | ENGAGE

Let us bring technology news to you.

<http://computingnow.computer.org>
Subscribe to our daily newsfeed

This article was featured in

computing **now**

ACCESS | DISCOVER | ENGAGE

For access to more content from the IEEE Computer Society,
see computingnow.computer.org.



IEEE  computer society

Top articles, podcasts, and more.



computingnow.computer.org