Extracting Meaning from Web Content

Michael Upshall, Consult MU

Michael@consultmu.co.uk

For many years, experts have attempted in vain to extract meaningful information and content from web pages. HTML is a dumbed-down language, holding almost nothing of any meaning (except for generic headings such as H1, H2 and so on, that will vary in meaning from one site to another); so any attempt to improve the communication of meaning from a web page is usually based around adding additional codes and vocabulary to HTML to make it more structured and meaningful (schema.org is an example).

Now Diffbot claims to be able to identify site content better. According to a post in Marketing Land (perhaps not the most reliable source), Diffbot has gained $10m in funding (quite likely) because it is "creating semantic Web content – that is, information that is characterised by its meaning – even though the page hasn't been formatted in that way." This sounds very unlikely, in fact it sounds like magic.

Certainly Diffbot does some basic things with a Web page. It separates a Web page into pictures, title and story – all of which HTML does already. The magic is Diffbot's other initiative: the creation of a "Global Index", a collection of knowledge (or graph database, a rather fashionable term, although there isn't a graph anywhere in sight) that will be searchable. Their goal is "to categorise most of the business-valuable information on the Web".

The graph database is what is used to attempt to classify the web page, along the lines of if an article contains the terms "bridge", "trump" and "trick", then the chances are it is about bridge the card game rather than bridge the civil engineering structure. Most likely, although Diffbot don't reveal their exact methodology, the tool will be cleverer than that and will use what Amazon terms "statistically improbable phrases" – phrases that occur very often in a particular type of book, enabling Amazon to recommend other books (or content) like it, containing the same SIPs. Why statistically improbable? As one blog response points out, this means that, although the terms "magic" and "London" occur frequently in Harry Potter novels, other phrases such as "Hogwarts" and "Hermione Granger" are less likely to occur in non-Harry Potter titles.

All very creditable, but does it work? I tried putting one of my own blog posts through Diffbot's "Test Drive" page. It added two labels to my story: "publishing" and "library" – hardly rocket science. To be fair, it did recognise me as the author.

Why is Diffbot special? Well, they have a very impressive PR department! A glance at the headlines from the media suggests that the world outside believes Diffbot is something special, with comparisons to Google and Intel:

- TechCrunch: *Don't Read The Comments—Let Diffbot Analyse Them Instead*
- Xconomy: *Could a Little Startup Called Diffbot Be the Next Google?*
- GigaOm: *Diffbot Aims to Convert the Web Into One Big Database, One Page at a Time*
- VentureBeat: *DiffBot's New API Brilliantly Reveals What's Hiding Behind Any Link*
- Wired: *Diffbot helps machines to read web pages like humans*
- Wall Street Journal: *Investors Back Diffbot's 'Visual Learning Robot' for Web Content*

How successful is Diffbot?

There is little explanation on the site, but the achievements of Diffbot appear to be focused around a limited number of Web pages, such as product catalogues and retail sites. The nature of these sites makes it easier to extract some kind of understanding from them – it's not too difficult to work out from an online store Web page which are the products.

Diffbot makes claims for the success of its system by comparing their performance with other competitors (AlchemyAPI, Embed.ly, Goose, for example), which shows them having an improbably high F1 score of 0.94 (the F1 score measures precision - how accurate the results are, with recall – how well the tool finds the links. It is generally assumed that two different human indexers looking at the same material would score no more than around 0.8 on this measure, since humans disagree on what is a correct link). However, in the small print of this comparison it would appear that the quoted F1 score measures only the extraction of title and text from a web page. You could say if you were being uncharitable that any bot could do that.