Journey to the Centre of the Web

Martin Dodge
University College London

Introduction

Where is the centre of the World-Wide Web? How do the tens of millions of pages connect together to form the Web? How does the density of inter-weaving connections vary through Web space, are some parts tightly woven together, while other areas are poorly linked, isolated backwaters? In this article I will explore how we can measure and map the structure of the Web to help answer these questions.

Everyday, millions of people start-up their browsers and head out into the Web in search of all manner of different things, but they have no map of the structure of the Web to guide their journey. The designers and managers of Web sites, who are trying to attract visitors, do not know whether they have a prime location on the Web in relation to the centres of traffic. Improving our knowledge of the structure of the Web can help producing, searching and browsing the massively growing, chaotic space that is the World-Wide Web.

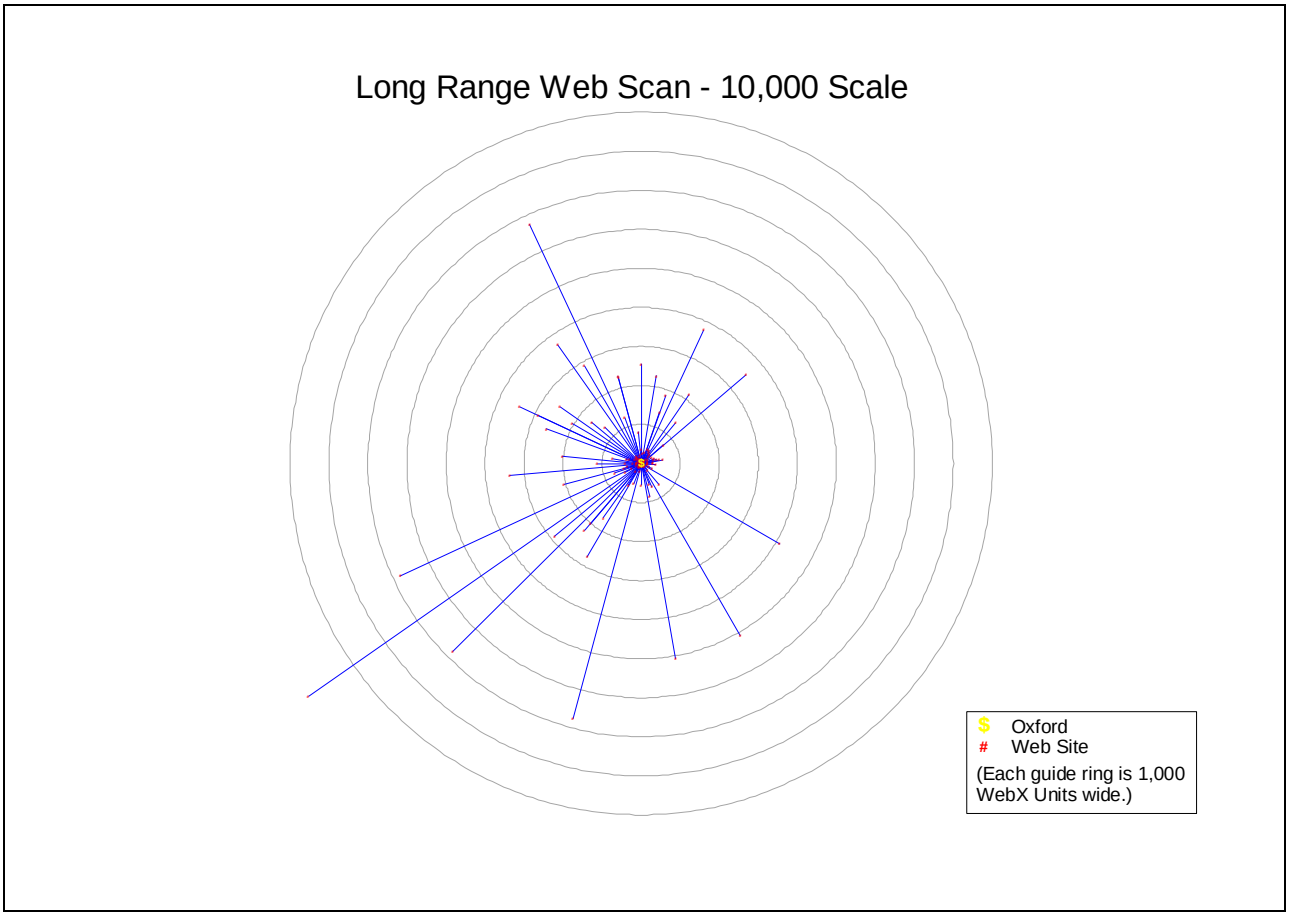
To analyse the structure of the Web I developed a metric of the centrality of Web sites called *WebX* (Web ConXections) *distance* based on the connections between sites**(1)**. This connectivity is measured by the outgoing and incoming hyperlinks. The structure of hyperlinks has been used by a number of academic researchers to analyse Web structure**(2)**, while an alternative method is to use Web traffic to determine the most popular and, therefore, central sites**(3)**.

To illustrate the potential of *WebX distance* metric I have analysed a subset of the Web, 122 sites of the universities and major colleges in the United Kingdom. I visualise the results as *Web Scans*, using an astronomic cartographic style.

To the Centre of the Web

Data on the size of Web sites of the UK universities and number of hyperlinks between them was gathered using the AltaVista search engine. A script was used to make the 14,884 separate queries to AltaVista, using the syntax `+url:<site1>.ac.uk +link:<site2>.ac.uk`, necessary to count the interconnection between sites. Over 450,000 links were reported, although the vast majority were internal ones within sites. The connectivity data was analysed to find the central Web site, that is the node that has greatest connectivity to all the other 121 universities. This was the Web site of the University of Oxford (`http://www.ox.ac.uk/`).

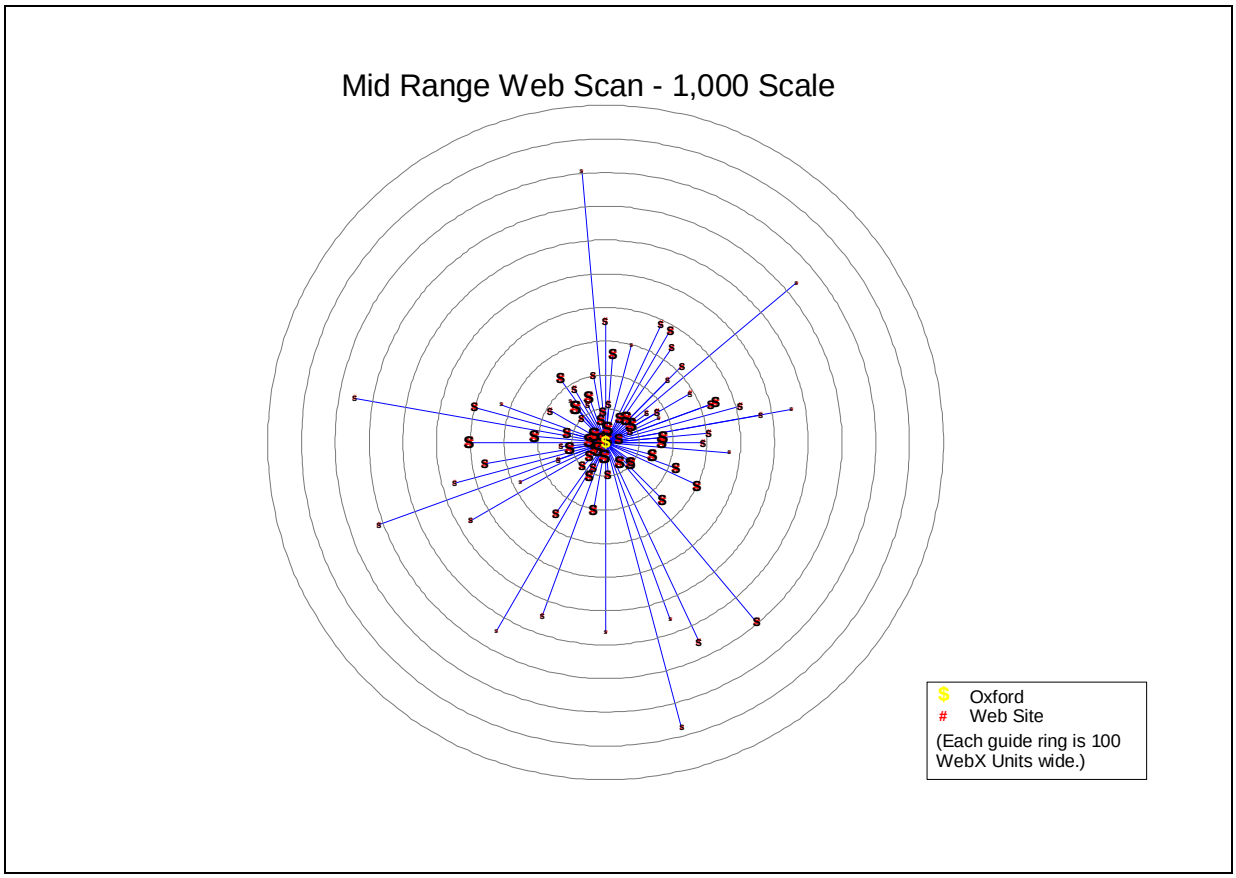
To calculate this Web hyperlinks were used to create a virtual distance measure - the *WebX distance* - which was inversely proportional to number of links between any two given sites. So the more links between two sites the closer they are in WebX distance terms. The WebX distance from each university to every other one was calculated and stored as a graph. This graph was analysed using a shortest path program to measure the distance from each Web site to every other one. The results were normalised by Web site size. Oxford's Web site had the lowest mean distance being the closest to all other sites. The centrality of other Web sites can then be expressed in term of their distance from the Oxford centre point. This measure is visualised in the following *Web scans*.



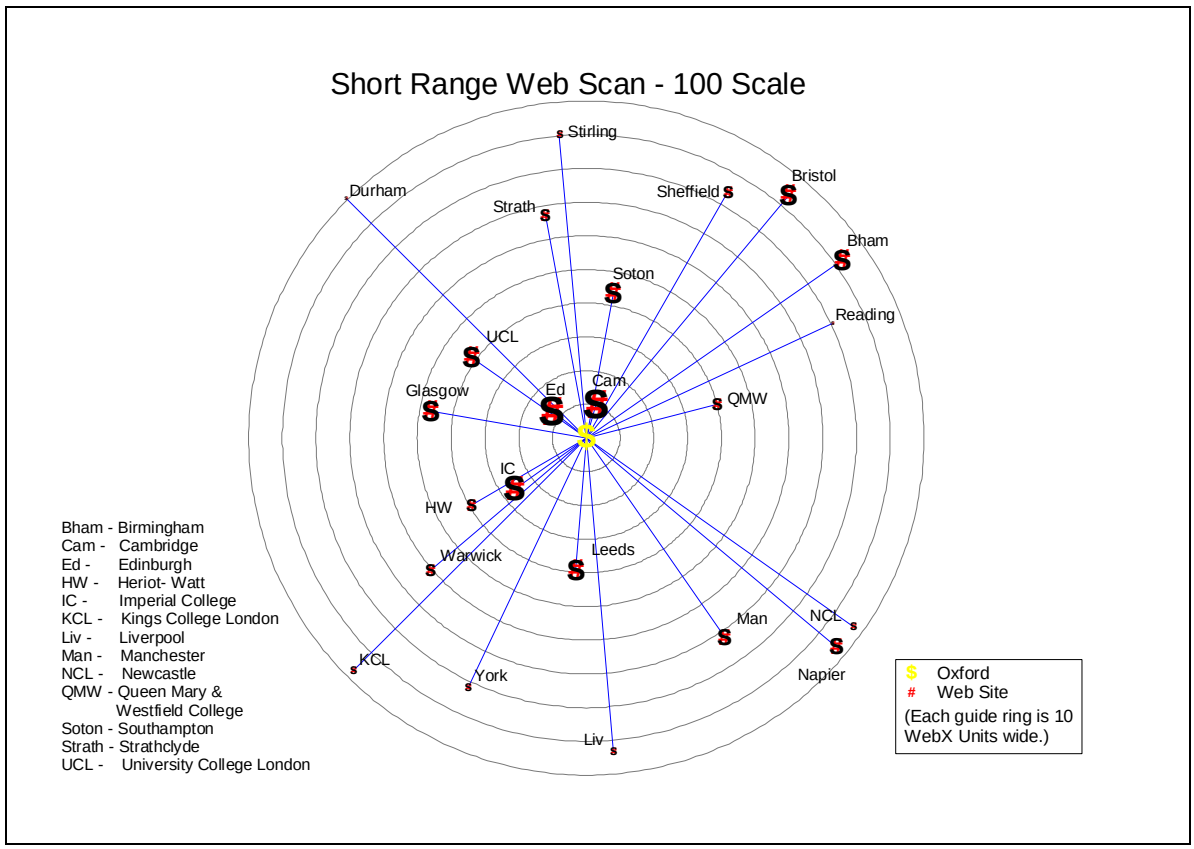
Web Scans

Taking an astronomic metaphor, the Web scans have Oxford's site as the sun. It is the central point of gravity and light for our small Web solar system, around which 121 planetary Web sites spin, their orbital distances being equal to their WebX distance from Oxford. Blue trace lines highlight this distance from the sun. The closer a site is to the Oxford sun, the more central it is. The orientation of the Web sites is arbitrary, based on five degree interval determined alphabetically from the name of the university, although in the future this position could be used to encode other data such as their approximate geographic location. Ring are overlaid on the scans, spaced at regular WebX distance intervals, to act as a visual guide to the scale of the system. Three Web Scans are shown in this article, mapping the system at different scales. They were created using a GIS (geographical information system), even though GIS technology is usually employed to analyse the real world, it is quite capable of mapping virtual worlds like the Web.

The largest scale Web scan, shown above, enables one to see the whole system, including the most distant Web sites. The furthest away - the equivalent of Pluto in our Web solar system - is the University of Wales Institute Cardiff (`http://www.uwic.ac.uk/`) way out beyond the 10,000 unit ring, by far the least well connected Web site. The most striking feature is the dense cluster of Web sites at the centre of this system, forming a dense red nebula inside the smallest ring. There are a number of peripheral sites, such as the University of Westminster, Humberside University, the London Business School and Dartington College of Arts. To see more detail of the heart of the system we switch to a mid range Web scan.



The scan above shows only the sites that are 1,000 units or closer to Oxford. At this scale the Web sites have become recognisable as planet-like disks rather than just red dots, their size proportional to number of pages the site contains. At 874 WebX units from Oxford, Liverpool John Moores University is the furthest out in this scan, followed by University College Chester orbiting at 807 and because it is a small site, with just over 700 pages, it is represented by a very small red dot.



This macro-level scan shows the 24 Web sites that are closest to the centre, these are the core of the academic Web in the United Kingdom. What is immediately striking are the two giant sites very close to the Oxford sun, the universities of Cambridge and Edinburgh, which have large Web sites and are very closely interconnected. Cambridge has a WebX distance of 10 and Edinburgh is only slightly further out at 13. Imperial College comes next with a WebX distance of 26, double that of the second place site. Further out from the top three, there is cluster of site around the 40 WebX mark. These are Heriot-Watt University, the universities of Leeds, Glasgow, Southampton, Queen Mary & Westfield College and my own institution University College London (UCL). All these universities are well connected, centrally-placed in the academic Web. UCL has a WebX score of 42, putting it in seventh place away from Oxford, a respectable position given its historic role in the development of Cyberspace in the UK, as it was the first organisation in Britain connected to ARPANET, the Internet's forerunner, back in 1973. Finally, right on the edge of this scan is the University of Durham's site with a WebX score of exactly one hundred.

Why Measure the Web?

Understanding the Web's structure, using metrics like my WebX distance, is important for a number of reasons. On one level the Web is worthy of analysis and mapping because it is the terrain of the Information Age. Just like surveyors map the real-world, so we should begin to chart virtual spaces. The Web is a unique and fascinating phenomena, an expression of human knowledge, creativity and desire to communicate. It is exhibiting great organic growth and therefore requires serious quantitative measurement to begin understand it. On a more prosaic economic level, the Web is widely predicted to become a significant source of wealth and employment, yet we know really very little about it.

The WebX distance metric as a quantitative measure of Web site centrality has potential practical benefit to those who use the Web as either users searching and browsing or as content producers. In terms of searching, research has shown that use of measures of centrality based on hyperlinks can improve the accuracy of the results from search engines**(4)**. Hyperlinks can be thought of as virtual citations, which can aid in assessing the credibility and veracity of Web sites. Exploring the Web by browsing from page to page, it is easy to chance upon an interesting looking page, but at present there is no help available to assess the value or reliability of the page. Knowledge of the page's connectivity and centrality would useful in this regard. Perhaps what is required is a dynamic Web scan type map built into the browser software so one can assess the location of the current Web page in relation to centre of the Web, it would enable users to see at a glance whether they are in the bustling downtown or lost out in the backwaters. The Web scan could also help find potentially useful pages and sites that are nearby but are not be immediately visible from your current location. The browser tool from Alexa Internet (`http://www.alexa.com/`) offers some of this functionality, being able to recommend sites similar to the one you are viewing, but lacks any map presentation.

For the designers, writers and maintainers of Web sites, metrics like WebX can provide a useful measurement of how well their site is performing. Building up connections and getting a site in towards the centre of the Web are important in raising visibility and generating traffic. Being able to determine the WebX distance of a site and also of your competitors would aid Web masters.

Web Galaxies in the WWW Universe

There are really many different centres of the Web depending on what you are interested in and looking for. We might envision the Web looking like the universe with many different galaxies, clusters and nebulas. There would be massive bright galaxies of new information in one sector, dark nebulas of long abandoned sites and nascent clouds of newly forming Web spaces. Different galaxies will be formed by Web pages and sites that have similar content (e.g. all movie sites) or that are related by country or language. Each of these galaxies will have their own central points and will

merge into larger cosmic structures. To map this one would need a dynamic Web scan that could visualise, at the appropriate scale, the part of the Web that one is interested in. I analysed one small Web galaxy, the sites of UK academia, which is a homogenous and well connected part of the Web Universe. The UK academic Web is part of large stellar structures, being an important component of the UK Web and also related to other academic sites in other countries.

At the heart of the ever-expanding Web Universe there are several extremely bright supernovae These super-centres are known as portals in the current jargon and are the key points where many people begin their Web explorations. There is currently a scramble by the largest sites to establish themselves as Web supernovae because of the promise of future financial rewards. It is not clear how many there will be, but some of the front runners at the moment are Yahoo, Netscape, Excite and Microsoft. It will be interesting to see, as we rapidly approach the new millennium, which of these portal supernovae go the way of real stars and burn-out and which remain shining brightly to guide our journeys through the Web universe.

Martin Dodge is a researcher at University College London and runs the Cyber-Geography Research project (http://www.cybergeography.org/). He can be contacted at m.dodge@ucl.ac.uk.

Notes

(1) I am very grateful to Naru Shiode for his assistance in calculating the WebX distance metric.

(2) See for example, paper by Tim Bray, "Measuring the Web" (http://www5conf.inria.fr/fich_html/papers/P9/Overview.html) and Jon Kleinberg's, "Authoritative sources in a hyperlinked environment" (http://simon.cs.cornell.edu/home/kleinber/auth.pdf.)

(3) Several commercial companies survey Web site popularity, including RelevantKnowledge (http://www.relevantknowledge.com/), Media Metrix (http://www.mediametrix.com/), Web21 (http://www.100hot.com/).

(4) See technical papers by - Massimo Marchiori, "The Quest for Correct Information on the Web: Hyper Search Engines", http://decweb.ethz.ch/WWW6/Technical/Paper222/Paper222.html; Jeromy Carriere and Rick Kazman, "WebQuery: Searching and Visualizing the Web through Connectivity", http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html. Also, try Google (http://google.stanford.edu/), an experimental search engine created by researchers at Stanford University.