The shape of the internet, inside and outside the corporate firewall

This is a review of the paper "Searching the workplace web", with some comments about how tagging may change the conclusions of the original research and challenging search engine engineers to look at the differences between inside and outside the firewall.

I have been discussing the efficacy of our internal search tools and how hard it is to find stuff, and to be honest, assumed that it was the crapness that most users accuse their IT colleagues of. However a colleague, Bernard Horan recommended that I read "*Searching the Workplace Web*", which suggests a different answer.

Searching the Workplace Web argues that intranet's are different from the internet and that more flexible and different search algorithms are required to search an intranet; the most successful internet search algorithms are not necessarily going to work well on an intranet.

The author's made four observations.

The first assumption is that content in the intranet is often created for the purposes of dissemination of an authorised opinion or fact, or a statement of policy. There is no design intent to attract readership. One observation on which this is based is the fact that often content in the intranet is very light of additional hyperlinks to suggest further reading or to quote sources. Why suggest further reading when your authoring policy? There is no further reading to be done. Why quote a source when the answer is "because my boss said so!" For example, check out your companies expenses or travel policy. The authors argue that a corollary of this is that "in list" based algorithms such as PageRank may be less effective on intranet searches. Interestingly, I ran this past Chris Gerhard, who said that he'd been looking for the original text of the Road Traffic Act, but that Google (an in-list based search engine) had difficulty finding it as it preferred commentaries on the law, because they were more referenced by web page authors.

These two examples take us to assumption two. In the intranet, we're often not looking for the "Wisdom of Crowds"; there are often very small result sets for a given query, often the correct result set is only one entry. This will occur when you are looking for a policy, or an officer's represented opinion. Will expenses pay this journey cost? Is this a supported configuration? It occurs when the researcher is looking for authority not opinion.

Observation three, is that there is (likely) to be less spam inside the firewall. (I wonder if this is an aging observation, with the growth of blogs and the opening of mail archives to search, it may be that this is weakening in strength, but its unlikely (but not unheard of) that large porn collections will be found be accident in an intranet search). The corollary to this observation is that some ranking algorithms that are unsafe on the Internet, become useful inside the firewall.

Observation four is that Intranets are less friendly to search. The authors observe that much content is held inside databases, or document servers, portals, directories and other specialised interfaces

While reading Benkler's "Wealth of Networks", I first came across the concept of, a shape of the internet. Obviously we all know that some sites are very influential and highly read, but the internet's hyperlinks have a topography that can be described and measured using graph theory. This was as far as I can tell first explored by Broder, Kumar, Maghoul and others in their paper "Graph Structure in the Web". These topographies were discovered during the ascent of the dynamic search engine, which won out to the detriment of the directory based references. These two papers are contemporaries and it'd be interesting to see if these topographies remain useful as insight today.

IBM discovered their intranet topology was different to the Internet, with a smaller "core" and a larger periphery. The core is a bunch of sites that meet formal graph theory definition as strongly connected. (See Graph Structure in the Web). The size of the OUT segment, pages that can be reached from the core, but do not return is larger than in the internet, and is much exacerbated by domino document repositories. The periphery is also much larger than in the internet, they can be found from the crawl seed pages (which must be in the IN segment) but not from the core. They measured the frequency distribution of the probability that an in-list based sort algorithm would place on a page's relevance and discovered a difference in shape between the intranet and internet results, with a lower proportion of high scoring pages in the intranet.

Another interesting innovation was that the research team created three indices (most solutions used only one) for determining relevance, these were content, title and anchor text. (Anchor text is the text between the anchor tags, and thus chosen by the author to represent the link in the original document). They then build a flexible ranking engine that had a number of input parameters. (I might write about this another day, but if you want more now go to the original document).

It's three years later and its almost certain that with the changes in user content authoring tools and the fact that there is more spam and more opinion, that the topology will have changed. The improved content creation tools represented by blogs and wikis also weakens the assumption that intranet content has low link counts. Sun is very permissive about blogs, as are as far as I can tell IBM, but the introduction of blog and wiki technology has both strengthened and weakened the firewall and hence the boundary to the intranet. Company staff are better informed and make better judgments whether to publish their material internally or publicly and can do so more easily both politically and technically, but the fact that sometimes/often the authoritative statement by a colleague is on a public blog, means that intranet search needs to pass through the firewall and "join" intranet and internet resources.

One very obvious example, illustrating the difference in intranet content can be discovered by examining tag clouds. If one were to compare my del.icio.us tag cloud with my internal delirious tag cloud, there are huge differences, I have no picture gallery inside Sun, none of my food, gardening and culture bookmarks are stored, internally I have a bunch of "How to" and "Do not do", repository links and applications home pages. (Let's face it a lot of the Technical documentation is on the internet now, and the secret R&D stuff I don't get to see anyway!) Also the clouds have very different shape, partly because I have over 1200 bookmarks on my public site and considerably less on internally. Tag clouds may also be another way of overcoming some of the four observations and corollaries

In order to "see" my true tag cloud, I need to "add" my private and public bookmark lists, which I organise using delirious & del.icio.us. It needs a form of federation.

It would be hoped that tags might be part of the answer, but the different shape of the intranet, may make the development of discriminating tags very hard. My experience at the moment is that this is true. I had given up, but I have been inspired to have another go.

Sites like Digg and del.icio.us where user generated content creates huge numbers of hyperlinks because of the number of users, will also distort the shape of the internet; as they become part of the core , the size of the "out degree" segment will become larger. Hyperlinks become the votes of web readers, not authors. Although its possible that since these sites are designed to be read by people, that there will be a more limited reference to them on other sites and they will remain part of the "In" segment. The XML feed services will however be referenced by many sites, and the linkroll gadgets mean that they are referenced..

So intranet search queries require a different approach to internet search, but is it getting closer or traveling in different directions.

tags: technology search internet intranet graph internet topology ce2.0