Corporate Search: Can We Just Get Google?

Arkadi Kosmynin, Jessica Chapman

December 2012

Put the two words “intranet search” in the Google search box and what do you get? In 2010, at the time of writing the first version of this article, the top link was titled, “Why intranet search fails: Gerry McGovern”. A text excerpt from this influential article read,

> *“Most intranet search delivers lamentably poor results. Time and time again, I hear staff plead: "Why can't we just get Google?”*

In his 2006 article McGovern argued that poor intranet search results are primarily caused by the poor quality of the contents. We disagree with this and will explain why below. However, during 25+ years experience in IT, we had never met anyone who is happy with their corporate search engine. That is, until very recently. Now the first link in Australia is titled “Arch Intranet Search Engine”!

So, why can’t we just get Google? It would seem that this is the technology that delivers excellent results on the global Web’s tens of billions of pages, so why should there be any problems doing the same on intranets that range from a few hundreds to a few millions of pages – tiny, compared to the size of WWW? Unfortunately, it’s not that simple.

Google won the search engines battle of the 1990s mainly due to two factors: first, it uses the anchor texts of web links to improve the descriptions of documents the links point to, including those that Google has not even visited, thus significantly increasing their index coverage. Even more importantly, Google uses a document “quality” component to compute the scores of search results. What this means is that Google shows the documents it finds in an order that is based not only on how well the documents match the query, but also on the quality of the documents. The effect is that the user sees higher quality documents first. How important is this? It was not obvious in the 1990’s that search engines should do this. They did not - and most of the pre-Google search engines are extinct now.

The method used by Google for quality estimation (the PageRank algorithm) is based on the web links graph. Put simplistically, the more web links there are that point to a web page, the higher is the estimated quality (rank) of that page. As we see, the methods that make Google work so well depend on the availability and nature of web links. These methods don’t work well for internal websites. The reason for this is the lack of statistical information required by the PageRank and similar algorithms to estimate document quality. While global web links can be used to estimate the popularity of a page that is viewed externally, intranet links usually reflect the site structure. For corporate sites, multiple links to the same page create redundancies that make the website harder to support, and web developers tend to avoid making these when possible. As one example, the CSIRO ATNF website (www.atnf.csiro.au) has a home page for a software package known as MIRIAD. This has 26 external web links pointing to it (significantly more than to most of other ATNF pages), but only two internal links. Based on a count of intranet links, this would not be detected as an important page.

Augmenting documents descriptions using anchor texts is also difficult when there are very few links. In the example above, Google has 26 text fragments to augment the description of the MIRIAD page. An intranet search engine has only two.

If the methods that Google uses on the global Web do not work well for intranets, is there anything that works better? Google Appliance (a corporate search solution) advertises a new feature: Self-Learning Scorer - a method for improving relevance judgements based on learning from user clicks on search results. Conceptually, this is not new. In 1997, Arkadi Kosmynin suggested this approach in an IEEE Communications Magazine article to be used in the global environment, not intranets. The problem is that it may take a long time for intranet search engines to "learn", since they receive far fewer hits than a global search engine. Fortunately, intranets have a much better source for estimation of document quality. This source is web server logs.

Normally, web servers maintain log files where they record details of every request that they process. These logs are instrumental for various tasks, such as analysing web site use patterns and problems. It is not hard to use these logs to count requests to each document for a certain time interval and estimate relative document quality, assigning higher quality scores to documents that are retrieved more often. This is simple, logical and reliable.

This method is used in Arch, an open source search engine based on Apache Nutch. Arch extends it by adding features critical for corporate environments, such as authentication and document level security. Most importantly, Arch replaces the Nutch document quality computation module with its own that uses web logs to estimate relative document quality.

Results? We asked a few people who know our site content well (approx. 70,000 pages) to evaluate the performance of Arch versus Google (the global engine, not the Google Appliance) in a series of blind tests on the public part of the site and versus Nutch on the whole site. It was the first time that we have seen people happy with an intranet search engine. One of the testers said,

> *"I've found so much good stuff that I did not even know existed. I was tempted to spend hours reading it instead of continuing testing."*

This effect is easy to explain. Arch ranks higher documents that people use more often, thus helping searchers to transparently discover popular resources. This also makes Arch a very effective tool for people new to the site.

We measured the "precision at top ten" documents: the number of correct hits returned by each engine in the top ten documents. Users had to mark correct hits, but did not know which engine returned which hits. Arch performed on average as well as Google and about 30-40% better than Nutch in our tests. The tuning and evaluation software that was used to do the tests is released with Arch. We encourage others to use the test software to evaluate search engines against each other and publish results. Hopefully, if there are enough tests done and results published, this will make the intranet search engine market more transparent.

What's next? Arch is available to the public free of charge and the source code is included. If you are happy with your intranet search engine, please let us know what you are using. We would love to know. If you are not happy with your current search engine, ask your IT guys to try Arch. It will make a difference!