An Enterprise Search Engine in 15 Minutes?

Arkadi Kosmynin, Jessica Chapman

It is not disputed that search engines are extremely useful tools. After all, who does not use global web search engines, such as Google, these days? Their smaller sisters, intranet or enterprise search engines, perform similar functions on a smaller scale and help share knowledge within the enterprise with an increase in staff productivity.

However, there are a couple of serious problems with most enterprise searches. Firstly and perhaps surprisingly, the algorithms that make Google so effective on the global web, do not work nearly as well on intranet websites. The Google search algorithms, used for the global web, are based on statistics obtained from web links. These statistics are insufficient even in large intranets. Finding effective algorithms for intranets and enterprise websites is highly challenging. We wrote about this problem in our article "Intranet Search: Can We Just Get Google"? If this problem looks interesting, do read it.

A second problem is cost. The huge cost of running global web search engines is almost always paid by advertisers. The cost of licensing, deployment and maintenance of enterprise search engines has to be covered, and terms such as "enterprise strength solution" or "enterprise license" often indicate high costs. For example, if you want a solution for your enterprise search needs, you may wish to purchase the Google Search Appliance, with prices starting around $32,000 for indexing up to 500,000 documents. Furthermore, this type of solution may not adequately address the first problem outlined above.

CSIRO Arch is an open source, free intranet search engine based on the freely available Apache Nutch. Arch modifies Nutch for use on intranets and extends it, adding critical features needed for corporate environments, such as document level security and access control. The result is a very scalable and versatile high performance, enterprise strength search engine that is capable of indexing intranets of any size. Best of all, Arch provides excellent search results for both intranet and enterprise searches. In systematic 'blind' tests, we found that Arch's performance on intranets is comparable to Google's performance on the global web – something that, as far we know, other intranet search engines have yet to achieve.

The recently released version 1.7 of Arch is very straightforward to install. In this article we explain how to deploy Arch in 15 minutes. You will need a computer with Linux/Unix or Windows Vista/7 + Cygwin, a couple GB of RAM, Java 7 or later version, Apache Ant and Ivy. .

To get started, go to the Arch home page http://www.atnf.csiro.au/computing/software/arch/ and download the latest version of Arch. Save it somewhere on your hard drive. Switch to the directory where you've saved the downloaded Arch package (suppose it is arch-1.7-src.tar.gz) and do the following:

```
#$> tar -xzf arch-1.7-src.tar.gz

#$> cd arch-1.7

#$> ant

#$> cd ArchHome/bin

#$> vi arch

#$> ./arch
```

When you type vi arch (see above), the Vi editor will open an Arch crawling script. Go through this and edit the parameters. As a minimum, you must provide seed URLs to start crawling with. It is also a good idea to do a trial crawl first, with crawling.depth = 2.

That's it! This can be done in a minute or two, so, why did we ask for 15 minutes? Because Ant has to download a fairly large number of dependencies (Java libraries) that Arch uses. Depending on your internet connection speed, this may take about 10 minutes. When Arch finishes crawling, you can query it at http://<arch host address>:8993/arch/search.

This quick setup will work well for simple small to medium intranets (from a few hundred pages to a few thousand). More complex cases may require a more advanced setup, configuration and tuning process.

For example, you may want to:

- exclude certain branches of your websites from indexing;

- hide restricted access documents from people who are not authorised to see them;

- prune documents before indexing, to prevent common parts, such as menus, footers and headers from polluting your index;

- split your websites into different areas for differential crawling and easier search;

- deploy a number of gateways for filtered search;

- dramatically improve the search quality by making web server logs available to Arch for statistical processing;

- make available for indexing and search exports from other (non-web) sources, such as corporate databases.

All these can be done in Arch, but will take longer to do. The Arch Quick Start Guide provides references to relevant chapters for the Deployment Manual. So, the answer to the title question of this article, “Enterprise search engine in 15 minutes?” is “yes and no”. “Yes” – if you have a relatively simple and/or small site to index. “No” – if you need to customise Arch configuration to suit your special needs.

The Pareto principle (also known as the 80–20 rule, the law of the vital few, and the principle of factor sparsity) states that, for many events, roughly 80% of the effects come from 20% of the causes (source: Wikipedia). One of the known application of this rule to the software area states that 80% of users use 20% of software features. An obvious outcome of this is that a well designed software system should make it possible, or even better, easy, for users to use just 20% of it’s features. We believe that Arch 1.7 does exactly that! If you are looking for an enterprise search engine solution, please invest 15 minutes into trying Arch and if you can think of something we could make even easier in Arch, let us know.