Google Search Appliance vs Arch – a Free Open Source Alternative

Arkadi Kosmynin

Enterprise search engines are an important tool for increasing staff productivity. A good enterprise search engine may save hours by helping to quickly locate information that is critical for decision making. One of the enterprise search market leaders, Google, is discontinuing their Search Appliance product over the next three years (2017-2019), replacing it with a cloud based solution. This may not be an acceptable transition for everyone, as some organisations would have to change their security policies to allow keeping sensitive data by a third party outside of corporate network.

In this article, I compare GSA with the enterprise search engine Arch that we developed at CSIRO and license under a CSIRO Open Source software licence, and argue that Arch can be a good replacement for GSA. This comparison covers the essential criteria that influence the cost of the solution vs its usefulness, i.e. value.

- **Scalability:** both Arch and GSA can work on clusters of computers and offer unlimited scale. The difference is in the price you pay. For example, Arch can index 500K documents on a single non-dedicated hardware box that costs less than $2K. To index 500K documents with GSA, you would have to pay $32K – just for one node.

- **Cost of deployment and maintenance:** both are easy to deploy and maintain, and offer almost a "turnkey" solution in simple cases. We discussed this topic in article "Enterprise Search Engine in 15 Minutes?"

- **Query power:** GSA supports wildcard searches, spelling correction and ordering on a set of document attributes. Arch offers customizable faceting out of the box together with full power of Apache Solr and very powerful Lucene query syntax that includes wildcard, proximity, fuzzy and range searches, boolean operators, term boosting and field grouping.

- **Supported types of indexed documents:** both GSA and Arch offer a set of parsers that cover all common document formats. For parsing documents, Arch uses Apache Tika toolkit that "extracts metadata and text from over a thousand different file types", is open source and can be extended to parse files in a proprietary format if needed.

- **Supported types of document sources:** Both GSA and Arch are able to index non-web data, such as the contents of relational databases. Arch uses Apache Solr as its index server. Apache ManifoldCF is a connector framework providing Solr connectors that let Arch index data residing in enterprise data repositories, such as FileNet P8, Documentum, LiveLink, Meridio, Windows Shares, SharePoint, relational databases and others.

- **Index completeness:** With web log processing enabled, Arch is able to provide a more complete index than GSA by finding isolated web pages that "normal" web crawling algorithms, including those used by GSA, will not find. Arch detects these pages in web server logs and will index pages even if they are not linked to crawling seed URLs. Arch can also be configured to work in "watch mode", where it checks new log information every few minutes and indexes new pages almost instantly when they are created.

- **Security:** both products support document level access control. Arch also supports an unlimited number of secure search gateways that can serve pre-filtered queries to narrow search for security or relevance reasons.

- **Flexibility:** both products have clearly defined APIs and extension points, but, being an open source software package, Arch is more modifiable, extendable, and therefore more flexible, able to accommodate virtually any custom requirements. Arch uses, and derives from power and flexibility of well known, widely used and actively supported open

source Apache products, such as Nutch, Solr, Tika, Lucene and Hadoop. This also ensures that expertise is available when a custom solution is needed.

- **Relevance of results:** arguably, this is the most important criterion that makes a difference between success and failure of the search, and success and failure of the search engine. Users want to find things that they are looking for, and preferably, on the first page of the results set. Achieving a good relevance on an intranet is not simple, because the algorithms that work so well for Google on the web, don't work as well in intranet environments. We discussed the reasons for this in the article "Corporate Search: Can We Just Get Google?" Arch solves this problem by using web server logs information to estimate document quality, which is a very important component in search results ranking. In a comparison of the performance of Arch and GSA on a real life document collection of over 100,000 documents, we measured the "precision at top ten" documents: the number of correct hits returned by each engine in the top ten documents. On a set of 47 test queries, Arch over performed GSA by about 10% on average. It appears that the performance of Arch on intranets is at least as good as that of GSA.

It looks like Arch and GSA are comparable by the criteria addressed above. However, being open source and thus more flexible, Arch may provide a solution in some cases where GSA options are limited. As Arch is free, flexible, and provides at least comparable to GSA performance in relevance, the most important quality criterion for a search engine, it clearly represents much better value for money for most use cases.