A Review of Search Solutions

By Jorge Handl - Software Architect - Flaptor, Inc.

The Value of Good Search

Every content-centric site needs good search. Without it, users won't be able to find what they're looking for, the site will loose the chance to satisfy their needs and revenue will be lost.

To create value with search and maximize revenue requires significant customization. However, tailoring most search solutions takes a lot of time and effort, or simply can't be done. Many sites either scrape along with a bad search experience, sink a lot of money into building an in-house custom solution with ongoing maintenance, or give up entirely and hand out their user experience to Google.

This white paper reviews the current search solutions, their pros and cons, and introduces a new product that addresses these problems.

The Hidden (and not-so-hidden) Costs of Standard Solutions

There is a wide range of search solutions, each with its own strengths and shortcomings. To analyze them, it is best to group them into three main categories.

1. Database solutions

Most modern relational database engines offer full text search capabilities among their features. These are very basic search systems that offer little or no flexibility and suffer from performance and quality issues when dealing with even moderately large numbers of documents. Relational databases are designed to deal with a very different problem, and full text search is not a good fit for that technology.

It is common for businesses to start with these kinds of solutions as they are already available in their database software and can be quickly implemented. But later on, when search starts to be a problem, they tend to look for more adequate options.

2. Custom Search Libraries

When faced with the need to add search to a website or to replace an inefficient database search facility, developers are often tempted to go the "build" route, based on the assumption that internal development is less expensive, can be better integrated and gives more control over performance and scalability, and on the availability of open-source search tools like Lucene, Solr and Sphinx.

An in-house solution has to be built, tested, debugged, optimized, hosted, maintained and updated. Those developers are diverted from working on the core business goals. In the meantime, the market does not wait and customers that expect a functional search are turned away or receive sub-par service, adversely impacting a company's revenue and reputation.

Building a highly scalable and high performance service on top of open-source libraries is not trivial and takes highly trained engineers with years of experience. However inexpensive a custom solution may seem, the true costs tend to be hidden and even then custom solutions fail to meet their intended user objectives.

3. Hosted Services

There are clear advantages to using a hosted search service. You don't have to hire experts, build infrastructure, develop your own search system and wait for it to be working. You're leveraging the expertise of those who specialize in search solutions, you're taking advantage of an infrastructure that has been already built and designed to scale, you're using a product that has already been proven to work, and you're riding the wave of innovation as you immediately use new features as they get available. You get all of that from day one at a fraction of the cost.

Who is providing these services? Clearly, Google dominates the scene, with Yahoo as a second runner. They both surely know how to provide search. Let's examine their offers:

Google Custom Search: Although this ad-supported solution is easy to use, it has many shortcomings. You don't control what is indexed or how long it takes to see a change, so the search results are often stale or incomplete. More importantly, you don't control the search results; Google decides what the best content to show your customers is, not you. Moreover, the ads are not optional and are chosen to increase Google's advertising revenue, so you may end up serving sponsored links for rival companies.

Google Site Search: This is the pay version of the previous service with some added perks. It comes without ads, you can make sure all your content gets indexed and you have a little control over search results: you can bias the results in favor of newer pages or a section of your site. Prices start low but only for very small numbers of documents, at $100/year for up to 1000 pages, and quickly grows to $2,000/year for 25,000 to 100,000 pages.

Google Commerce Search: Specifically designed for retail sites, this service gives the customer a lot of control through custom ranking of search results, promotion, faceting, filters and auto-complete. It comes at a steep price, though, starting at $25,000/year.

Yahoo! BOSS: Similar to Google Custom Search, it provides a window into the search giant's web search index, it also lacks control over what is indexed, when it is updated and which results show up for a search query. Their service is currently ad supported but they have been giving mixed signals on whether or when they will start charging for it, having announced an upcoming fee-based structure in February 2009 and again in August 2010.

IndexTank

After years of providing consulting services for companies with search needs through the use of open-source search tools and even producing their own open-source search tool (hounder.org), Flaptor saw a market opportunity in providing hosted search, leveraging their extensive expertise in the field.

In July of 2010 IndexTank was born and took as its first customer the social bookmarking site Reddit.com, solving what had been a thorn in Reddit's side for a long time.

Their Solr-based search was plagued by problems: it either returned no results because of server overload or returned mostly irrelevant results. Users seldom found what they were looking for. Posts asking why search wasn't getting fixed were frequent enough that the admins wrote a FAQ page citing lack of resources to throw at the problem. Several alternative Reddit search sites were built by Reddit's users out of frustration. When IndexTank took over, users loved it. They even contributed by writing a reference manual for it.

IndexTank's features are comparable to Google Commerce Search ($25k/yr) at a fraction of the price: $300/year for up to 50,000 documents and $1,200/year for up to 300,000 documents.

IndexTank offers complete control of search results ranking through free-form functions, document promotion for specific queries, faceting, query auto-completion and search filters. IndexTank goes further with exclusive features like geolocation support, real-time indexing and real-time update of scoring attributes.

As the Reddit experience proves, IndexTank is flexible and fast enough to handle millions of documents, hundreds of queries per minute, and can rank search results based on real-time data such as user up- and down- votes.

IndexTank
CUSTOM SEARCH YOU CONTROL

A Review of Search Solutions

Jorge Handl
Software Architect
Flaptor, Inc.

The Showdown

Here's the pricing and feature set provided by the hosted solutions already discussed:

	Yahoo! BOSS	Google Custom Search	Google Site Search	Google Commerce Search	IndexTank
Ranking Control	No	No	Date & section bias	Limited rules	Free-form functions
Promotion	No	No	No	**Yes**	**Yes**
Faceting	No	No	No	**Yes**	**Yes**
Filters	No	No	No	**Yes**	**Yes**
Auto-complete	No	No	No	**Yes**	**Yes**
Geolocation support	No	No	No	No	**Yes**
Real-time indexing	No	No	No	No	**Yes**
Real-time attribute updates	No	No	No	No	**Yes**
Analytics	No	No	No	Yes	Coming Soon
Indexed object	Web Page	Web Page	Web Page	Product data	Arbitrary document
Price for 50,000 docs	**ads**	**ads**	**$2,000 / year**	**$25,000 / year**	**$300 / year**