Deep Web TECHNOLOGIES

Quality, Not Quantity

The danger of overlooking quality of search results

If the goal of federated search is to facilitate the rapid location of the most relevant information, then one would expect customers of the technology to place quality of results at the top of their features list.

Why quality of results matters
When libraries, research organizations, and IT departments pursue federated search solutions they usually focus on the *features* of the various offerings. Most will purchase the solution which offers the features they have selected at the lowest price. Frequently overlooked is the importance of *quality results*. Yet, in the final analysis, impressive features cannot overcome the problems caused by poor search results. If the goal of federated search is to facilitate the rapid location of the most relevant information, then one would expect customers of the technology to place quality of results at the top of their features list.

Most don't.

What does "quality of results" mean, anyway?
Objectively measuring quality of search results is very difficult. All of us have had the experience of looking at a set of search results and being able to instantly determine that the top few results are relevant or not. While it may be hard to define quality of results, we know it when we see it. The best results based on our search terms rise to the top.

Too many results, not enough time
With the exponential growth of information comes the greater-than-ever need to filter that information since none of us have the time to assimilate all of it, or to do the filtering ourselves. Google prides itself on delivering millions of results for many of its queries. But, in this overwhelmed state we're constantly in, "more" is too much. If you've ever used Google to perform scientific research, did you find it useful that Google identified a million hits for your query? Were the top few results spectacular? Did you find yourself having to evaluate the results and the sources who provided them one by one? Did you find this time-consuming? The bottom line is this: federated search needs to pick the right documents from the huge set that's competing for your attention.

The dirty little secret of the search engine industry is that Google doesn't give you the results that are most relevant to you, just the ones that enough people consider worth linking to.

It's not a popularity contest: the dirty little secret of the search engine industry

There are a number of fundamental differences between the way that federated search engines find the best results and the way that Google and the other popular search "crawlers" do. The dirty little secret of the search engine industry is that Google doesn't give you the results that are most relevant to you, just the ones that enough people consider worth linking to. But, "popular" and "relevant to you" are not the same. High quality scientific and technical documents may not be popular, yet some may be vitally important to you. Conversely, popular papers may not be relevant to you.

The need for speed and the price you pay

A major objection to federated search, mostly from casual users, is that it's not fast enough. Why wait 30 seconds for search results when Google will give them to you nearly instantly? The short answer is that Google doesn't find the quality documents that federated search engines do. Google follows links from one web page to another, gathering up documents it finds along the way. It produces an index of all the documents which allows it to match queries to documents very quickly. Federated search engines dive into "the deep web," where the best research material exists. The cost of searching the deep web is time. It can take some number of seconds to access a deep web source and retrieve its best content. However, you pay a huge price for getting a million hits in a hundredth of a second (what you miss when all you search is what Google finds). From this perspective, waiting 30 seconds to save a few hours is a good trade off

The myopic focus on features

If speed isn't worth the cost and if federated search gets you the best documents, then where do product features fit in? Features are important to the extent that they support the core functionality of a federated search solution. Surely it's important that a federated search solution provide ease-of-use, results management, alerting, and other capabilities that are up to par, but if your users aren't happy with the top results they're getting, then the features don't mean much. Surprisingly, most vendors focus myopically on features and sell features while ignoring quality of results as the most important benefit.

What really matters

What follows is a condensed list of the most important guidelines in choosing a federated search solution. Not surprisingly, all of the factors directly affect the quality of search results.

1. **Choose sources wisely.** Your federated search solution is only as good as its sources. Choose quality, vetted sources that are diverse and relevant to you. In selecting a vendor, make sure that they can "connect" to all of your important sources. A connector, simply put, is the code that the vendor software uses to retrieve what you are looking for from a particular information source. Don't assume that any vendor can provide any connector.

The very best crafted connectors perform advanced (fielded) search much better than poorly crafted ones.

2. **You want craftsmanship.** What few vendors will tell you is that quality of search results is directly impacted by the quality (the workmanship) of the connectors. All connectors are not the same. The very best crafted connectors perform advanced (fielded) search much better than poorly crafted ones. There are many tricks to building outstanding connectors and the proof is in the quality of the results they get.

3. **Go back to the well.** Another secret is that many sources return a limited number of documents per query. Federated search engines that are smart enough to go back to a source for more results can potentially get more relevant documents. Not all vendors use this approach because it takes more work to code the connectors to retrieve more documents and it takes more work for the federated search engine to implement the logic to know when to "go back to the well."

Many people are unaware that many content sources rank their documents poorly, if they rank them all.

4. **Don't take their word for it.** Many people are unaware that many content sources rank their documents poorly, or not at all. A smart federated search engine will determine its own relevance for documents returned by each source, regardless of the importance the source gives to its documents. Having the federated search engine perform the relevance ranking is also critical because it's the only way for a user to be able to compare a document from one source to that of another and to know which is more relevant.

5. **Get the best ones to the top.** All relevance ranking formulas are not the same. The challenge of federated search is to sort documents from multiple sources given just small bits of information -- typically title, author, and perhaps an abstract or a snippet from the article. The better federated search engines can do more with this small amount of information, especially if they employ the previously discussed techniques to get the right set of candidate documents to begin with before they even consider ranking them. Note that some federated search applications don't even aggregate results from among the different sources.

That makes it very difficult to find the most relevant documents from the collective results set.

6. **Looks *do* matter.** Once you have the best set of results for a query, sorted in the right order, presentation of results is important. When evaluating a vendor, ask yourself if the results are presented in a way that makes them easy to view and navigate. If visualization or categorization of results matters to you, does the vendor offer clustering or other means of intuitively organizing or drilling down into your results?

Federated search is the right technology for research organizations looking to streamline their knowledge discovery processes.

How Federated Search fits the bill

Federated search is the right technology for research organizations looking to streamline their knowledge discovery processes. With federated search you get these important benefits:

- By picking the sources and letting the federated search engine scour them for the most relevant information you have the peace of mind that you've left no stone unturned.
- By selecting a diverse set of sources you stay current in your field while keeping a pulse on developments in related areas.
- By choosing a vendor who crafts smart connectors you can be confident that your federated search solution will squeeze the best information from each source.
- By getting the right features to augment precision searching your users will be certain to have a productive research experience.

Not all federated search engines are created equal

Deep Web Technologies is obsessed with getting you the best possible search results. We examine each source to discover all of its search potential, and then build our connectors to take advantage of their capabilities. Test drive one of our publicly available search portals and compare results to your other favorite search engines.

Here are a few of the research sites we've built:

- Science.gov
- Mednar.com
- Scitopia.org
- ScienceResearch.com

See us on the Web: http://www.deepwebtech.com