Northern Light

Northern Light Perspectives

Integrated Search vs. Federated Search

In a strategic research portal, the means by which the system aggregates research content and the search engine presents search results is a critical matter as it determines the utility of the solution and, ultimately, the business value of the research content that the user organization has developed and/or purchased. Experience demonstrates that, for this application, **integrated search** is a distinctly superior approach to **federated search**.

With **federated search**, the search engine sends the user query to the search engines native at each source, receives back a search result from each source (in, say, an XML package), munges the various search results together in some arbitrary fashion such as "first responder wins" and then presents them to the end user.

The attractions of federated search include the fact that federated searching obviates the need for a local repository on the enterprise network to index. And there are federated search connector vendors that will license a company the code required to put a syndicated source into a federated search solution. Also, Microsoft is supporting federated search in the ubiquitous SharePoint Search and has published a standard that all search engines at all content providers can adopt to join a SharePoint-based federated search application. Since the XML package of results provides a URL to invoke the document, users that are authorized by their network or cookied for the content source should be able to click through to the document from the search result.

There are however, several critical flaws to federated search for strategic research portals. These include:

- Indexing strategies vary by content source. Some may provide proximity searching; others may not, so proximity searching is not practical in the portal research application. Some may provide case-sensitive search; others may not, so case-sensitive searching is not practical in the research portal solution. Some content sources may provide unstemmed as well as stemmed indexes (more on this later); others may not, so literal searching is not practical in the research portal solution. The list goes on and on. Federated search devolves to "lowest common denominator" searching from a search strategy perspective very quickly.

- Supported search syntax varies by content source. Indexing strategy and search syntax are closely related ideas. The purpose of syntax is often to give access to specialized indexes. Without a common indexing strategy for the content in the research portal, it cannot be searched with the same syntax. In addition, some sources may support

Boolean operators such as ‘not’ or complex Boolean expressions, while others may not. Or if they do, they may use different conventions for invoking a specific search function. Alerting can be very risky in federated search solutions because there is little visibility into how the various sources process queries. And a content sources supported syntax may change from time to time invalidating saved searches and alerts, like Google recently dropping the ‘+’ operator that meant ‘and’ to everyone that has used Web search since Alta Vista introduced the ‘+” operator in 1994. This also drives a lowest-common-denominator syntax with little more than simple ‘and’ and ‘or’ statements generally workable, if that.

- Taxonomies vary by content source. With federated search, taxonomies are the option of the content provider and no two will agree on structure, terms, extent, or any other facet of taxonomies. With federated search of disparate sources, the idea of a consistent taxonomy across all sources that reflects your research needs is unobtainable.

- Text analytics is impossible with federated search. In a similar manner to taxonomies, because the full-text of the documents is not available to the search engine in the research repository for indexing, there can be no text analytics applied to the documents searchable in the research portal. Text analytics operates by finding text strings and relationships between text strings that imply meaning. Without the full-text to operate on, this process cannot be carried out.

- Accurate relevance ranking is impossible with federated search. Each of the search engines at the individual sources returns search results in ranked order using that search engine’s relevance ranking method. However, the federated search results consolidator has no idea how any particular source’s search engine arrived at the rank order. Each content provider can decide what elements and factors to weight and whether to rank them high or low for relevance ranking purposes. If there are 30 content providers in a federated search, there are 30 uncoordinated and hidden relevance ranking strategies in play.

  Even worse, there is no rational basis for the process that munges the various search results from the different providers to determine how to interleave them for presentation to the user. How do you decide if hit number 3 from Source A is more or less relevant than hit number 30 from Source B or even hit number 300 from Source C? There is neither a statistical means nor information theory basis for making this determination. So federated search engines use a variety of chance methods to do the munging including “first responder wins” (ironically rewarding the source that processes the query the least), random interleaving, source-quality based interleaving, and re-ranking based on the small bit of XML-returned search summary text for each hit. This last method is the current most popular with federated search connector vendors and also the one that produces the worst user experience since it destroys the one piece of reliable information in the search results from the various providers: the within-source rank.

Northern Light has a completely different approach: **integrated search**. We aggregate all the content by obtaining a full-text original copy of every document from every source. Then we index it with the *Northern Light Search Engine* using perfectly consistent indexing, taxonomy, and text analytic strategies uniformly across all sources. This produces an *integrated index* that can behave the same way across all sources in terms of search syntax and relevance ranking. Using Northern Light's integrated approach:

- Indexing strategies are consistent across all sources. All sources have the same set of indexes such as stemmed, unstemmed, proximity, case-sensitive, and acronym.

- Search syntax is consistent across all sources. All sources can be searched using the same, advanced, flexible, and powerful search syntax. Alerts can also use this same consistent syntax and be depended on to work the same way for all sources. There is no risk that saved searches and alerts will be disabled for one or more sources by changes to supported syntax at a content source.

- Taxonomies can be applied across all sources. With Northern Light's integrated search strategy, a client can use a relevant taxonomy applied to all sources in the research portal. Control of taxonomies resides entirely with the client, not with the content sources.

- Text analytics is enabled across all sources. Because the search engine has access to the full-text of the documents, powerful text analytics like Northern Light's *MI Analyst*™ are possible. Without text analytics, research portals are not using the most powerful tools available for speeding the time to insight.

- Relevance ranking is accurate across all sources. With integrated search, Northern Light can in fact tell if hit number 3 from source A is more or less relevant than hit number 30 from Source B or even hit number 300 from Source C. The relevance ranking formulas are applied to all documents in a consistent manner and every aspect of the document can contribute in the same way to the document's rank in the same way regardless of its source. There is no *munging* with integrated search, only *ranking*.

# # #

(12/15)