Enterprise Search Evaluation Guide

Seven critical characteristics your enterprise search solution must have

Enterprise Search is no different to any other IT category – vendors have different names for similar features or benefits of their products. This is often confusing for customers. We think that in the end it all comes down to 7 critical factors customers must evaluate when implementing an Enterprise Search solution.

These seven critical characteristics are: Relevancy, End-User Experience, Reach, Freshness, Access Controls, Scalability and Total-Cost of Ownership. The first four characteristics have a direct effect on the usage and usefulness of the search solution for the end users. The last three are important from a business point of view. In the following sections we explore each one of these characteristics: what they are, why they are important and how they can be measured. At the end we also provide a simple enterprise search evaluation worksheet that you can use to refer back to the top 7 factors that matter in enterprise search.

Relevancy

The purpose of an enterprise search solution is to help users find information. In simple terms, relevancy measures the "quality" of the results a search system returns, and how quickly users can find the exact piece of information they are looking for. While there's not a single standard measure of relevancy, most experts focus on the precision of the search solution. Precision is defined as the percentage of queries that are satisfactorily answered by the top "N" results the search returns. Often, precision is evaluated at the top 3, 5 and 10 results.

Precision is easily measured through a simple manual process. Search quality evaluation starts by collecting queries that were issued to the search system and the results that were returned to the user. Each result is normally inspected by human assessors and judged for its relevance to the query. A result will typically be considered relevant if it is a good answer to the query and satisfies the information needs of the user who issued the query.

It's easy to see how a search engine that returns the desired piece of information at the top of the results is more relevant than the one that requires the user to scroll to the third page. Even differences within the first 10 results matter to the end-user.

But that's only half of the story. A good search solution will return relevant results across all kinds of queries – for every user, every time without the need of human intervention. A search solution must maintain its relevancy even as new content and types of content are added to the system. Search solutions that require extensive and constant changes to the algorithm and tagging of content (what is commonly known as "tuning") cannot guarantee they will constantly produce relevant results as content grows. In short, tuning is the manual process where by a search administrator assigns different weights to different words at indexing or search time, adjusts the algorithm of the search system, or modifies the way queries are processed to make them appear higher or lower in the search index.

End-User Experience

End-user experience is about making it easy to find information. The bottom-line is that the ultimate proof of a search solution is its utilization – increased usage signals increased value for the end user. A combination of different factors will determine how good is the experience the end user has with the search solution.

- **Speed.** How quickly the search engine returns a user's query is critical in determining usage and satisfaction. Most enterprise search queries should return the initial result set in a second or less.
- **Readability.** Once the search engine has returned a result, the user must determine which one of those results is the most useful to them – and a good search engine won't make him scroll from the top of the page. The information that a search engine provides along with each result helps the user determine this. For example, grouping is a technique used to improve readability of the results: similar documents are collapsed into one entry to assist the user in making his choice faster. Query term highlighting helps the user find the query terms within the larger document.
- **Query Assistance.** A good search engine will not only return relevant results for a given query, but it will also help the user determine whether he made the "right" query. Automatic spell check is something that is easy to understand. Advanced features that help the user at query time include automated query expansion, context-sensitive stemming and synonyms, and suggested queries. Stemming and synonyms is the technical name of the feature that allows a user looking for information on "public park" to find those documents including "public parks" and "city parks", but not those including "public parking".

Reach

Reach of an enterprise search solution refers to the types of information it can provide search over, and thus, find for the user. Users have come to have certain expectations for search engines, and those include being able to find every single piece of content "out there". Reach can be measured along three dimensions:

- **The type of content it can provide results for.** Content can be classified in structured and unstructured content. Structured content refers to that one found in databases and line of business applications. Although different business applications provide some level of search functionality, the problem is that users must log-in to the right business application to find that content – which seriously limits the usefulness of those search capabilities. Unstructured content refers to documents, spreadsheets, presentations and other files that don't have a pre-defined or standard schema. Enterprise content, information and knowledge is increasingly located in unstructured content.
- **The places it can reach the content in.** While structured content primarily lives in databases, unstructured content can be found in file systems, web sites, file shares and other places. It's important that the enterprise search solution you deploy can get to all the places where your enterprise content lives.
- **Formats it can support.** Format is something that is more relevant for unstructured data. Regardless where the content is located, it is important that the search engine is able to open and index it – independently of what program was used to create the file.

Freshness

Content and information is always changing. A search system is practically useless – and potentially dangerous- if it returns outdated information. Before deploying a search engine, it is important to understand how it ensures the content it returns is fresh. There is of course a balance to fresh content – in order to keep its index fresh, a search engine must continually crawl the content sources. Unnecessary crawls can adversely affect the performance of the source systems. On the other hand, infrequent crawling leads to outdated information. The search solution you deploy must be intelligent enough to adjust itself to the different repositories and crawl them as efficiently as possible to keep the information fresh. In addition, the search solution should be able to accept requests from the systems where the content lives that indicate (based on changes or events) when the content needs to be updated or re-crawled.

Access Control

Organizations have content that is available to anyone behind the firewall, as well as content that is limited to a subset of users. An enterprise search solution must mirror the access control policies you have in place and only return those results that the user is allowed to access. A secure search engine should be thought of along three axes: real-time, use of existing controls and fine-grained, document-level security.

- **Real-time security.** Search engines are able to produce quick results for users because they keep an index of all the available information, which they can process and optimize to produce relevant and fast results. But the security policies or access control lists (ACL) in your content sources change frequently. This could potentially create a problem – you have to be certain that as permissions on the content change, they are reflected on the users searching for information, so only the people allowed to access the information are able to get results including that information when performing a query. Therefore, for maximum security the search solution should perform access control checks in real-time, rather than relying on outdated caching or manually configured lists.
- **Document-level security.** If you are enabling enterprise search in unstructured information, you must make sure that your search engine is capable of enforcing access permissions at the document level and not only at the directory level. Similarly, with structured information, you must make sure the search engine can enforce access permissions at the row-level. For example, if you're enabling real-time search in a CRM system, it's likely that certain users will only have access to a limited set of customers accounts – so it's important your search engine gives users access to only those customers accounts they're privileged to see, and not to all customer accounts.
- **Leverage existing controls.** Finally, but equally important, the search engine you deploy must work with your existing security systems. There are two reasons for this. The first is that in order to provide secure search, the search engine must be able to communicate with and understand the security policies that you have already implemented. This includes any Single Sign-On (SSO) systems that you may have deployed. The second reason is that the value of the search service declines, from the users' point of view, if they're required to have an additional password to search for information. Finally, if a search solution requires you to create an additional set of access control policies and permissions, the overhead becomes unmanageable and you in turn put your organizations' information at risk.

Scalability

Scalability of an enterprise search solution refers to three key metrics: the number of documents the search engine indexes, the number of queries per second that it can serve to concurrent users and the amount of man-hours it takes to administer the solution as it grows.

Before you select an enterprise search engine, it's important that you understand the amount of content you have (number of documents you have across all your content sources). Your enterprise search project might start relatively small, but it will undoubtedly scale up to include more sources of content and cope with the growth of digital information in the enterprise. So it's important that in addition to knowing how many documents you have today, you also estimate how many documents you will accumulate in the next few years. Most customers plan with a 3-5 year timeline in mind, and analysts estimate that content is growing in the enterprise at a rate of 40-70% CAGR. You don't want to go through the cost and effort of deploying a search engine that will only be useful for the number of documents you have today and not scale to your future needs. Similarly, you also need to evaluate the load the search engine will face – how many queries do you expect to have at any given time, and what's the experience that you want to provide your users with.

The key point you should keep in mind when thinking about scalability is that although number of documents and queries per second sound like two very different measures, they are indeed highly correlated, and it's impossible to evaluate one without the other. It's worthless to be able to index 10 million documents if the performance of the search engine slows to a "crawl". A good enterprise search solution will maintain its ability to serve query results even as its document counts grow.

It's impossible to separate the scalability conversation from the cost one. As you think about scalability, you should also ask yourself what kind of additional hardware investment it will take from you to meet those document count and queries per second you'll need to provide in the future.

Finally, the search solution should also scale in terms of the manpower required to administer it as it grows. It is important to highlight the fact that a good search solution will require as little time as possible to be up and running and will require even less on-going maintenance and administration effort. A search solution that maintains its relevancy as new content is added and doesn't require any additional maintenance work or "tuning" is far more scalable that a search solution that requires ongoing maintenance work from an army or search administrators.

Total Cost of Ownership

Of course, no enterprise information technology evaluation would be complete without looking at its total cost of ownership. When evaluating different enterprise search solutions, be sure to accurately budget for the following items: software license cost, hardware purchase cost, implementation costs, yearly software license maintenance fee, yearly hardware maintenance costs, ongoing search administration costs, user training, etc. Also, be sure to have accurate, benchmarked data from vendors about how much hardware is required to support the scale (document counts and queries per second) that you need both now and in the future. All too often, vendors quote minimum system configurations and real deployment costs skyrocket.

Enterprise Search Evaluation Worksheet

Seven critical characteristics your enterprise search solution must have

	Vendor 1	Vendor 2	Vendor 3
Relevancy			
Percentage of "answers" in the top __ results			
End-user perception about relevancy			
Amount of "tuning" needed, hours/week			
End-User Experience			
Average time, in seconds, it takes search engine to return results			
Search page readability: search term highlighting, cached version of document, relevant summary, grouping, etc.			
Query assistance: spell-check, automated query expansion, query stemming, synonyms.			
Reach			
Is it able to index both structured and unstructured content?			
File systems & databases it supports			
File formats it can support			
Freshness			
How often will it crawl content sources			
Is it able to search structured content in real-time?			
How will it determine how frequently to crawl sources?			
Does it accept requests for update/recrawl from the content source?			
Access Control			
Does it provide real-time security?			
Does it provide document/row-level security?			
Security systems it works with			
Scalability			
Maximum number of documents it will be able to index with initial hardware configuration			
Queries per second it will handle with initial document count			
Queries per second it will handle assuming maximum document count			
Amount of additional "Search Administrator" time required as content is added.			
TCO			
Software license cost			
Hardware purchase cost			
Implementation cost			
Yearly software licenses maintenance fee			
Yearly hardware maintenance costs			
Search administration costs			
User training			