white paper

Rocket

How Enterprise Search Enhances Enterprise Intelligence

How Enterprise Search Enhances Enterprise Intelligence

Revised July 2015

How Enterprise Search Enhances Enterprise Intelligence

While it's a myth that humans only use a small fraction of their intelligence, it's a reality for organizations looking to put more of their information assets to work. Here's how they can.

Never before have organizations been able to retrieve so much information so fast — both internally from their file servers, databases, and mail servers — plus externally from the Web. Yet, most organizations gain relatively little utility from all the relevant information they can retrieve. Of course, it's possible to find almost anything if you know where to look or what to ask for. However, raw access is rarely the problem. More often the reason organizations don't actually use more of the information they have is:

a) Users don't know what to ask for
b) Users don't know where to look
c) Users don't know the relevance of the information that they do find

When searching for information, what's important goes well beyond the rudimentary functions many people associate with search, like query expansions (such as compensating for spelling errors). We search not just to find information but also to overcome the barrier of not knowing what we don't know. Using information is difficult if we aren't aware ahead of time of what information we should be looking for. More than just *having* information, intelligence includes knowing *what* information matters and *why* it matters.

The not-knowing-what-you-don't-know problem also impacts organizations looking for search tools. They are less likely to find better tools if they don't know how much more advanced their search tools could be. For example, their frame of reference might not include cross-silo search capabilities. They're accustomed to searching unstructured text (one silo) with a search engine like Google and searching structured databases (another silo) with a relational database system like Oracle. So it might not occur to them that a single tool exists that can search both silos in the same search.

Their frame of reference might also not include natural language processing. They may think that the only way to find information relevant to a search is to know beforehand what search terms to specify. So, again, it may not occur to them that a tool exists that can discover relevant information when all the relevant search terms are not known beforehand.

In this case, the advanced capabilities to look for would include stemming, synonym expansion, classification, and other natural language processing algorithms. When combined, these functions reveal related information based on meaning rather than on just keywords. If organizations don't know what to look for in a search tool, like cross-silo search or natural language processing (or the benefits of those features), then they are less likely to find them. Similarly, they are less likely to find the information they need in order to accomplish the tasks for which they sought a search tool in the first place. But the converse is also true. With more advanced search capabilities comes greater enterprise intelligence.

What's Wrong with Open Source?

Open source search solutions take a long time and require advanced technical skills to implement. In particular they are very difficult to tailor to a particular organization's search requirements. Open source solutions also don't come with a user interface, so one must be developed; either that or the open source solution must be somehow integrated to an existing solution (like a content management system) that does come with a user interface.

Deep Versus Wide — The Missed Intelligence Opportunity

If you want to know more about enterprise search, you can start by asking this question: Why not just use a text search engine or a relational database system to find the information you need to take action? After all, if you type almost any keyword into a search engine you'll likely get thousands of results within milliseconds. By the same token, if you enter a word to "find" in one of a database's search fields you'll retrieve all the records — out of potentially millions — in the database that contain that specific word in that particular field. So what's lacking in these types of searches?

In both cases, what you'll get is a very wide search versus one that is very deep. The search will be "wide" in terms of the number of items searched (and possibly found) but not very "deep" in terms of intelligence gained — i.e., in terms of learning new information you might want to know about these items. Until that additional information is revealed, you might not even know this is information you would have wished to find in the first place.

Consider unstructured text search. Yes, you may get thousands of pages of results, with the most relevant results listed first. But that still leaves scanning the titles and short descriptions to glean which of those results might actually contain relevant information items (entities) you think you need to know (extract) from search results. Those could include names, paragraphs, images, addresses, dates, customer comments, reviews, and so forth.

In addition, you might also wish to have items classified by various inferences, as in: "Which product reviews were positive?" That type of inference is called sentiment analysis. Another type of inference is entity co-referencing, which is to identify which items are related to each other. An example might be to infer that a merger was about to take place. Even if the searched text did not specifically mention the word "merger," the inference of a merger might flow from finding company names, dates, stock prices, the names of investment banks, etc. That inference might *not* be found in a simple unstructured text search without co-reference inferencing.

Neither are relational database management systems designed to support this kind of depth. Their typical use case is someone wishing to identify any records containing a specific word in a specific column. But that only helps if the person knows in advance which columns to search across. It is possible, however, that highly relevant information may be in the database but not explicitly labeled, or whose label might not be one that would occur to the user. A good example could be a paragraph in a document. Unless the paragraph occupies its own column within a database, it is not possible to narrow your query to matches

occurring within a specific paragraph. Even if the document did come up in a search, you would still have to read through it in order to find whatever parts might be relevant. And you might miss the document altogether since the document might mostly be about a subject other than the one you're interested in researching. Other "intelligence enhancing" aspects of search—including entity extraction, entity co-referencing, and sentiment analysis — are also outside the scope of most relational databases.

Another limitation is that relational database systems only search the data contained within the database itself. In other words, each search method is confined to its own data silo — relational database systems are for structured data while search engines can search across both structured and unstructured data such as file systems and Web pages. That limitation creates barriers to enhanced enterprise intelligence beyond just the inconvenience of having to use different systems to access different silos. It also requires knowing ahead of time where the data you are looking for is likely to reside, which is something that enterprise search should be telling you (i.e., it's another opportunity for enterprise search to enhance enterprise intelligence).

Search tools that can't bridge data silos also can't find relationships between data residing in separate silos. For example, an alternative energy technology company might wish to correlate the names of prospects in its database with various environmental organizations' websites and other online mentions. Or it might wish to identify "up and comers" in the environmental movement by correlating events posted in calendars and news stories listed online with dates, names, and locations in a database.

Not finding relevant relationships, just like not finding relevant entities, causes diminished enterprise intelligence. If the user already knew what to look for, why it's relevant, and where to look, they wouldn't need to look for an enterprise search tool. They could instead rely on the search that is integrated within each data repository. The reason they can't is because these tools leave blind spots, which keep growing right along with the ever-growing volume of enterprise information. The tools don't see across data silos, they can't infer relationships or sentiment, and they can't identify relevant entities like keywords and field labels without explicit human assistance.

The fact that enhanced enterprise intelligence requires so many capabilities in a single enterprise search solution brings to light another issue: how well they all play together. For example, are they a collection of stove piped functions, such as search engines and relational databases, that IT will be called upon to integrate? And how will the solution support large numbers of enterprise users who may need to access the same function at the same time?

Given the impact these issues have on both enterprise intelligence and performance, the search capabilities organizations might wish to consider fall into five key categories.

Questions You Probably Don't Need to Ask When Picking a Search Solution

Does the solution offer a customizable thesaurus (the ability to add or replace synonyms for a word)? Does the solution support stemming (the ability to search on all words with the same root, or stem, as the word specified)? How about query expansion (the ability to automatically reformulate a search query, such as to take into account spelling errors)? These and other basic search techniques have been around since the 1960s. More meaningful are questions around scalability, entity extraction, clustering, and the other modern enterprise search capabilities discussed here.

Five Key Enterprise Search Capabilities

These capabilities cover both what search functions are present and how they are implemented, as in:

1. **Are key text analytics functions present?**
   - *Entity extraction.* Identifying items in text that belong to predefined categories, such as the names of people, organizations, location, and monetary value. For example, in the sentence "Bill Smith acquired 51% of the outstanding shares of XYZ Inc."
     - Bill Smith would be identified as the name of a person
     - 51% would be identified as a percentage
     - XYZ Inc. would be identified as the names of an organization
   - *Entity co-reference resolution.* Deriving the correct interpretation of text by connecting pronouns to the right individuals. For example, Gary is an investor. He invests all the time. "He" should be connected to "Gary."
   - *Relationship Extraction.* Finding links between previously extracted named entities. For example, all entities including dates, locations, and people associated with the same meetings.
   - *Sentiment analysis.* Classifying text based on emotional tone, such as positive, negative, or neutral.
   - *Faceted Search.* Progressively narrow your search via guided navigation and category drill down. For example, all competitors who sell products in various categories at various price points.
   - *Clustering.* Finding all documents that are related in some way without necessarily knowing ahead of time why they are related. For example, all documents related to a class action lawsuit.
2. **Does the solution have a pipeline architecture?**

Putting so many search functions in the same tool raises the question of how they will be integrated. A pipeline, or plug & play, architecture means that the various functions can be added or subtracted without needing to re-program the solution. This offers several advantages over trying to implement all functions as a single monolithic unit:

- *Best of breed.* Not only do multiple functions need to coexist within a single solution, but also any function can potentially be implemented using any of hundreds of algorithms, with new algorithms being developed all the time. A pipeline architecture allows you to select algorithms based on which is best, rather than on which is part of a packaged set.
- *Investment protection.* Being able to swap out older algorithms for newer ones means that your investment always stays state-of-the-art.
- *Extensibility.* Plug & play means you can not only update existing functions with newer, better algorithms but you can also add entirely new functions as new ways to search are discovered, without having to buy a completely new solution.
- *Conserve cash.* If the solution is extensible you can start with a limited set of functionality and add more down the road.

3. Is the solution scalable?

If a solution is highly scalable it means that many users can employ the same function and the same algorithm at the same time without experiencing processing delays, and without the need to deploy multiple copies of the solution (or your data) on multiple machines.

4. Does the solution support open standards?

Enterprise search tools need to be interoperable with other technologies, including the Web, content management systems, relational databases, and various file systems. For example, does the solution support Content Management Interoperability Service (CMIS), the open standard that allows different systems to work together within an enterprise network?

5. How secure is the solution?

An enterprise-grade search solution will rigorously protect the organization's information on a "need-to-know" basis so the solution cannot be used as a back channel into those assets.

Benefits of Advanced Enterprise Search

When enterprise search is built this way the barriers created by information silos are removed, and unanticipated insights can now emerge that take into account information deep within and across those silos. No longer are users required to know what they don't know. They can find entities without knowing, in advance, the "right" search phrase or field name to specify. They can also discover which events, people, dates, locations, documents or other items are related — without already having that insight. And they can see which members of specified groups belong to certain other groups without the foresight to search on any particular member.

These are all examples of enhanced organizational intelligence, which has actually been available to the enterprise for a long time. A key reason it's been missing is that, ironically, users did not know what search capabilities to search for. But they do now.

5 Questions to Answer Before Implementing Enterprise Search

What do users want to find?

Are they looking for entities? Do they want to group entities into clusters? Are they looking to find related documents? Asking users what they'd like to know will tell you which search capabilities to implement within a search solution.

Where is the data?

You'll need to know the major data sources. The Web? A database? A file system? A content management system? You'll need a tool that can talk to each of the specified sources, hopefully without a lot of configuration work up front.

What are your datasets?

Once you know your data sources you'll need to know the specific targets to be searched within those sources. For example, if the user wants to access customer comments, websites, contracts, and board minutes, what are the websites, records, and documents to be searched?

How much data is there?

The more data there is to search, the more hardware and networking resources required, and the more scalable the search software needs to be.

How often is data updated and how often will users want to access it?

Users may not want to access data that is no longer current, so data refresh needs to be synced based on when users need to access it.

Rocket Software

Rocket Software provides Enterprise Search and Text Analytics solutions that help users find the most accurate, relevant content they need to make smart decisions. Our enterprise search platform incorporates sophisticated indexing and analytic engines, paired with powerful search capabilities, to deliver an exceptional user experience.

To learn more, visit us at: www.rocketsoftware.com/solutions/enterprise-search-and-text-analytics

white paper

Rocket

www.rocketsoftware.com
info@rocketsoftware.com