

Why Search Matters

Information Retrieval in the Pharmaceutical Industry

Eric June



ARTVILLE

Information retrieval (IR) deals with the representation, storage, and organization of unstructured data, applying natural-language processing, semantic relationships, linguistic analyses, behavioral histories, and “fuzzy” statistical techniques to help human beings quickly find and retrieve the information they seek.

Eric June is chief software architect at AssurX, Inc., 305 Vineyard Town Center, Suite 374, Morgan Hill, CA 95037, tel. 408.778.1376.

Pharmaceutical manufacturers are awash in a sea of data. Laboratory and manufacturing information, nonconformance data and corrective and preventive action (CAPA) plans, clinical trial reports, supplier audits, and correspondence with regulatory agencies, customers, and other stakeholders all flow into the ever increasing sea. When data meets the US Food and Drug Administration’s definition of an electronic record, the agency requires the record to be readily retrievable and available for review. Regulatory requirements aside, if an organization is unable to quickly retrieve relevant data, business crises may ensue. According to Susan Feldman, research vice-president for Content Technologies at IDC (Framingham, MA), “Decisions are usually information problems. If they are solved with poor or erroneous information, they put the life of the enterprise at stake. It behooves the enterprise to provide the best information finding tools available, and to ensure access to all its intellectual assets, no matter where they reside.”

The ever increasing sea is made up of both structured and unstructured data. Structured data has a predictable form and typically is stored within tables inside relational database management systems (RDBMSs), or in structured file formats (e.g., XML). The techniques for reliably locating and retrieving structured data, such as structured query language (SQL) statements, are well understood and commonplace. Software application vendors often include tools that empower end users to easily construct sophisticated queries. User inputs are automatically transformed into SQL, thereby shielding the user from the complexities of SQL construction, while still allowing the RDBMS to optimally process the user’s query.

Unfortunately, these techniques address only a small part of the problem, because structured data typically makes up less than 20% of the total data in an organization’s information stores. Most of the remainder consists of text-based, unstructured data: free-form notes, procedures, letters, reports, faxes, e-mail messages, and a plethora of similar records and documents. The sheer volume of unstructured data makes rapid location and retrieval more difficult. That difficulty is compounded by the fact that unstructured data is directly created and maintained by humans.

Humans create and share information via robust but necessarily less-structured means such as concepts, analogies, exam-

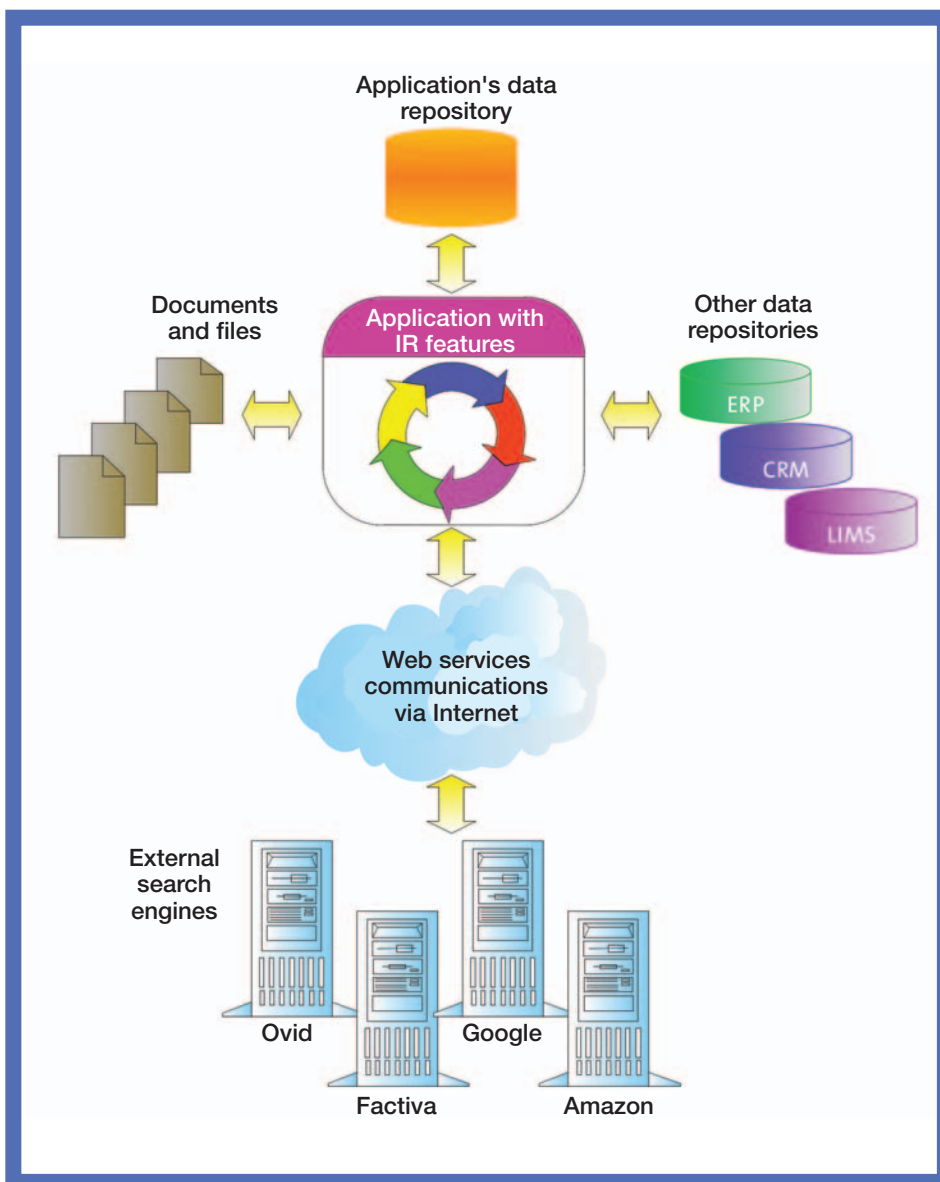


Figure 1: An enterprise-wide and global search architecture.

ples, and syllogisms. While structured data may have precise value constraints enforced by information systems during data entry, unstructured data derived from human communications generally cannot be rigidly validated via automated means. Unstructured data may contain abbreviations; typographic, spelling and grammatical errors; and variances based on regional dialects and individual style. For example, humans can tell that the phrase “the substance’s pharmacology” is equivalent to the phrase “the pharmacological properties and characteristics of the substance,” but machines designed to assess only structural factors may see no similarity at all.

Parting the sea

Information technologists have long recognized the limitations of SQL and related techniques when applied to unstructured data. The information retrieval (IR) discipline of computer science was created to research and develop more effective so-

lutions. IR deals with the representation, storage, and organization of unstructured data in general, and textual documents in particular. While SQL relies on a precise Boolean syntax for locating structured data using literal or syntactic attributes, IR techniques use natural language processing, semantic relationships, linguistic analyses, behavioral histories, and “fuzzy” statistical techniques. An immediate goal of IR is to assist humans in quickly finding and retrieving the information they seek. In a sense, IR techniques try to read users’ minds and determine what they really want, rather than demand that users state precisely what is on their minds. In the longer term, IR technology is destined to be a cornerstone of true artificial intelligence.

Software vendors serving the pharmaceutical industry are now infusing IR technology into their software applications. The feature set is commonly referred to as full text search, and most use familiar paradigms such as the search box and advanced search forms that have become ubiquitous on the World Wide Web. Specific implementations and their capabilities vary widely, however. Some allow only for simplistic keyword searches within the structured data maintained by the particular software application, while others allow sophisticated concept searches using

natural language syntax to be performed against structured and unstructured data that may exist anywhere within the enterprise.

These wide variances in information retrieval capabilities among different software applications lead to disparities in the ease of use and overall utility of the software. After all, software creates value by allowing people to retrieve and use data in new and innovative ways. As Susan Feldman points out, “while the costs of poor information finding are enormous, they are hidden within the enterprise and are therefore rarely perceived as having an impact on the bottom line.” Perception and reality are often very different. If information technology professionals are to maximize real bottom line benefits, they must first understand the capabilities of various IR techniques.

Seek and you shall find

Virtually all IR implementations allow users to search for literal words or phrases, and to look for any or all of the words

and phrases in the corresponding data source. These rudimentary capabilities are similar to traditional SQL-based query techniques for structured data in that one or more literal values may be sought, in one or more locations, via Boolean logic expressions.

A less common and more powerful technique is the concept search. Instead of searching for literal data values, concept searches look for terms with similar or related meanings. IR systems use thesauri or related word collections to process concept searches. For example, a concept search for “adverse” will also find its synonyms “detrimental,” “inimical,” and “deleterious” when a thesaurus search is specified. A related concept search for “string” will also find “violin.” Violin is not a synonym for string, but a conceptual relationship exists because violins cannot function without strings.

As mentioned above, humans occasionally introduce typographical or spelling errors into the unstructured data they create. Simple literal searches will be unsuccessful in locating such data, since the misspelled words are not literally equal to the properly spelled search terms. IR technology solves this problem by allowing for fuzzy searches, optimally with a user-defined level of fuzziness. For example, a search for “pharmacology” with a fuzziness level of 1 will find “pharmmacology” and “pharmacology” (in addition to the properly spelled “pharmacology”) because all of the misspelled terms have single-character typographical errors. Higher fuzziness levels will find even more derivatives.

Human foibles aren’t limited to typographical or spelling errors. Consider this introductory sentence in a hastily composed memorandum: “Their is an irrefutable conclusion supported by there study.” If the IR tool supports phonic (homonym) searches, “there” or “their” may be successfully located regardless of which term was actually used in the search.

Proximity searches allow words or phrases to be located when they occur near other words or phrases in unstructured data. For example, in the memorandum sentence above, the phrase “irrefutable conclusion” is within 4 words of “study.” This proximity relationship may be used in a search to find other study-derived irrefutable conclusions, while excluding conclusions that either weren’t irrefutable, or were irrefutably derived from other sources (e.g., internal experiments). Some IR implementations use a nonspecific definition of nearness, typically expressed in searches with the keyword “near.” More advanced IR implementations allow precise specification of nearness, such as the exact number of intervening words, and the order in which the words or phrases must occur (before, after, or in any order).

Variable term weighting provides the ability to emphasize particular words or phrases within a larger set of words or phrases in a search request. For example, it may be desirable to weight “irrefutable” five times more than “conclusion” to increase the chances of finding irrefutable information of all types, including conclusions, in the search results.



Figure 2: An excerpt of the results of an integrated search of multiple data repositories.

The top 10 attributes of a good information retrieval solution.

1. Global reach via open, standards-based extensibility to information inside and outside the enterprise.
2. Aggregates and sorts results from many sources.
3. Natural language concept and synonym searching via thesauri, related words collections, or both.
4. Regular expression searching.
5. Fuzzy searching to “see through” typographical or spelling errors.
6. Accommodates fielded data constraints within searches.
7. Stemming searches with international language support.
8. Proximity searches with proximity precisely definable.
9. Variable term weighting.
10. Phonic (homonym) searching.

Stemming is a technique used by IR tools that allows searches to automatically expand to include word derivatives. For example, a search for “apply” will also find “appliance,” “applied” and “applying.” Stemming rules are specific to a particular language (English in this example), so global pharmaceutical manufacturers should seek IR tools that include international language support. These more robust tools will also allow for selective processing of accent characters, case sensitivity, and other linguistic factors based on language or locale.

Regular expression searches enable arbitrarily complex textual patterns to be matched, and are especially applicable to the pharmaceutical industry. For example, organic and inorganic compounds are named systematically in accordance with standard conventions and nomenclatures. Regular expression searches may be used to find references to compounds that are related or similar to a particular compound by searching for patterns in the compound names. Regular expressions may also be used to reliably locate terms that may be expressed differently based on locale, or in modern *versus* Latinized forms (e.g., “hemoglobin” *versus* “haemoglobin”).

The newest IR tools augment their exemplary unstructured data searching capabilities by melding in techniques borrowed

from their structured data retrieval ancestors. A Boolean syntax similar to SQL may be used to limit searches to fielded data values (*i.e.*, data values that exist in particular locations). The “fields” may be explicit, such as the fields in relational database tables, or implicit and determined by inference via analysis of unstructured data during the indexing process (*e.g.*, a bolded and centered line near the top of a document that is inferred to be the title). For example, these capabilities allow a user to locate “adverse reaction” or similar concepts anywhere within RDBMS records, but only if the product name field contains one of three specific values.

Extending the reach

Deployments of enterprise-class software solutions for enterprise resource planning (ERP), customer relationship management (CRM), laboratory information management systems (LIMS), quality management systems (QMS), *etc.*, typically seek to eliminate silos of information through system integration. At the

same time, software vendors are beginning to introduce powerful IR features (full text search) into their enterprise-class software products. Since most software vendors limit the reach of their IR features to the data maintained by their own applications, new types of silos are starting to emerge: silos of searchable data (maintained by new applications) and silos of opaque, non-searchable data (maintained by legacy applications). Surely there has to be a better way.

One way to avoid creating these searchable-*versus*-not-searchable data silos is firmly in the hands of software vendors. Rather than introducing proprietary, closed IR solutions, they can choose to open their search tools to external data, files, and search engines via standards-based interfaces. The benefits of doing so fall squarely in the lap of their customer. While rolling out the next generation ERP, LIMS, or QMS system, the customer can coincidentally and with little extra effort roll out the next generation, enterprise-wide IR solution.

Figure 1 depicts a system of this type. The application’s full text search features span across its own data repository, external files, and enterprise data repositories that are maintained by other applications (*e.g.*, ERP, LIMS). Web services technology is used to add additional value by integrating external search engines such as those provided by Ovid, Factiva, Google, and Thomson Gale. Application users can now search across any or all of the available data sources, whether inside or outside the enterprise. The application coordinates the execution of the search, aggregates the results, and sorts the aggregated results based on user-defined criteria.

Figure 2 shows an excerpt of the results from a global search of this type. The user has searched for “Lactose” and specified the search scope to include:

- subsets of the application’s own data (issues, actions and file attachments);
- medical journal articles indexed by Ovid;
- internal documents (current revisions and archived prior revisions);
- books available from amazon.com;
- supplier data from the corporate ERP system.

The user chose to sort the aggregated results by relevancy.

Conclusion

Recent advances in information retrieval technology can help pharmaceutical manufacturers stem the tides of their ever-increasing seas of data. Decision making, problem solving, and regulatory compliance are all optimized when relevant information can be retrieved faster. As software vendors continue to infuse IR technology into their products, it is imperative that information technology professionals in the pharmaceutical industry understand the capabilities and limitations of various techniques if they are to fulfill their mission of providing strategic technology guidance to their organizations. Product and technology selections should be made with an eye on the future. Optimal search and retrieval solutions extend to data in any location, whether inside or outside the enterprise. **PT**