IDC
Analyze the Future

WHITEPAPER

The Agile Enterprise: Enabling Signal Based Processing Over Rule Based Systems

Carrie Solinger and David Schubmehl

Sponsored by Lucidworks

White Paper

The Agile Enterprise: Enabling Signal-Based Processing over Rule-Based Systems

Sponsored by: Lucidworks

Carrie Solinger David Schubmehl
November 2016

IDC OPINION

The increase in unstructured information combined with the increasing pressure to improve knowledge worker productivity has made locating the right information at the right time more imperative than ever. In many studies, IDC research has shown that workers are not happy with the search capabilities that their enterprise search systems provide. Information-driven organizations need to improve their capabilities around enterprise search in order to maximize revenue, manage costs, and increase productivity. A decade ago, researchers and information scientists thought that the best way to improve search could be achieved by increasing the relevance of the results returned by the search engine. Google and Yahoo! were pioneers in providing these kinds of improvements in the consumer arena, but this approach did not yield gains for the enterprise. To provide similar quality of results, search vendors developed the concept of rule bases and "directed search." These rule bases were how search administrators could programmatically direct users toward specific results for specific queries. While this approach worked well, it tended to require lots of administrator care and feeding and resulted in many custom rules (sometimes numbering in the hundreds).

Meanwhile, modern search has evolved dramatically over the past five years, especially within the area of web search. Leaders in the web search space, such as Google and Facebook, are now using contextual clues, machine learning, and user interaction data to provide contextual relevance and better overall search. Lucidworks Fusion with its new Signals capability brings this same sophistication to the enterprise. Signals allows search administrators to easily create customized relevant answers to search queries based on what the system believes is the most useful answer that the user is looking for.

Lucidworks Signals provides the same types of technologies and capabilities that modern artificial intelligence (AI) and cognitive systems use to enable users with sets of recommended answers and "next best actions" based on their actual usage of the system as well as the wisdom of crowds. The use of machine learning and contextual clue recognition and action greatly improves enterprise search capabilities over traditional rule-based approaches that have existed for the past decade. Organizations should strongly consider tools like Lucidworks Fusion and Signals to drive these types of improvements with their enterprise search systems.

November 2016, IDC #US41915716

IN THIS WHITE PAPER

This white paper provides information and guidance on the next generation of enterprise search tools using data-based analysis, contextual clues, and data-driven machine learning to provide accurate and relevant search results for organizations and their users. Executives who read this white paper will learn about the benefits and trade-offs of contextual and data-based analysis and recommendations and answers provided by the next generation of enterprise applications that use search-based features and capabilities to improve performance and add value.

SITUATION OVERVIEW

The world of search systems and technologies has changed dramatically in the past five years. Although organizations have used web search tools like Google to help them find information on the web, their internal enterprise search solutions have lagged far behind. These enterprise search systems are based on outmoded notions. Organizations are facing an unprecedented challenge from the overwhelming amounts of information, emails, reports, websites, PDFs, recordings, and videos that are available to them. Their information is located in multiple applications and repositories across multiple time zones and datacenters. In addition, some of the most valuable knowledge of an organization is locked in what IDC calls unstructured data. Inside all of this unstructured data are facts and insights about every aspect of an organization's business. Social media, blogs, and news stories often contain information and commentary about an enterprise's product marketing approaches and competition. The combination of increasing information across a wide range of repositories and the reliance on outmoded notions of relevance and recall has created a nightmare for most organizations. "Our search is terrible!" and "We can't find anything!" seem to be the rallying cry of today's knowledge workers.

However, it does not have to be this way. The availability of a wide variety of data along with the technology, skills, and processes to take advantage of this type of data has emerged over the past few years. This is promising to radically change how information is accessed, analyzed, and shared. Signal processing offers capabilities to make better decisions, personalize customer interactions, optimize operations, and innovate. A big part of realizing this promise is dependent on efficient and effective access to unstructured information — especially in context — and the analysis and organization of such data.

Role-specific search-based applications are becoming popular because the jobs of knowledge workers are specializing and their information needs are varied. Research scientists require access to general research journals as well as patent filings and the organization's own research information. Customer service agents need access to product documentation, order histories, and prior discussions with customers. In each case, these knowledge workers need access to a range of information, but the information type varies by job function. In both cases, the search-based applications developed for these roles make information available from several repositories and include functions and workflows specific to the knowledge worker's task. Organizations developing these role-based search applications also must tune their systems to return the information most relevant to their particular query. In the past, the tuning ability of most search systems has been rule based or what the search technology industry likes to call "guided navigation."

Rule-based or "guided navigation" systems for search have been around for the past 10–12 years. These systems are oriented toward heuristic-based approaches for determining search results, using a combination of hierarchical structures and deterministic rules. These rules look something like "use Chicago" if the search is "main office" or "head office." A rule-based system identifies certain keywords

that trigger the applicable rules. In many search systems today, there are literally hundreds of different rules for many different types of search situations. The advantages of a rule-based system are that it has a uniform structure, there is a complete separation of knowledge from the processing, and the rule-based system has the ability to deal with incomplete and uncertain knowledge through the use of heuristics. These rules are located in separate structures from search indexes and can be applied to multiple types and sets of indexes. In some ways, one can think of these as taxonomic structures used to help improve information discovery.

However, the disadvantages of a rule-based system for search are also significant. From what IDC has seen with these types of systems, the rules get outdated, the reasons these rules were put in place were based on assumptions that may no longer be true, and business conditions have changed. This is the biggest flaw in rule-based systems: They are inflexible and are unable to learn and adjust automatically as the content and the types of searches adjust and evolve over time. Keeping the rules fresh and up to date requires an administrator to review, edit, and adjust these rules and add new rules over time. Given that the average search installation in an enterprise has less than one full-time resource assigned to its care and feeding, this type of work usually goes undone, and over time, the rules get out of sync with the content in the indexes and the queries from users.

In contrast, a signal-based system uses search context and user interaction data to derive insights about what users are looking for and how best to provide it to them. Signal-based systems can react and adjust search responses and order them according to the relevance that is obtained from user analysis and context about what the user actually wants, not what some set of rules say the user should want.

A signal could be a click. It could be a query. It could be a document view. It could include which emails were opened. Typically, signals are events with time stamps that provide information relevant to search. For example, clickstream data provides a cascade of time-stamped data points, such as user A searched for term X, and then user A clicked on document Y and then on document Z. Raw signal data is generally a large set of small data points that in and of themselves are not informative without further processing. Aggregation is the "processing" part of signal processing. An aggregator reads in raw signals and returns interesting summaries, ranging from simple sums to sophisticated statistical functions.

These individual data points are collected, aggregated, analyzed, and then used to provide clues to the search system about what the user is looking for and trying to find. These clues are then used to provide the most relevant answers and/or recommendations that can be used to get the best answer or provide the "next best action." These types of systems can replace rule-based systems, making them obsolete. New systems, based on signal processing, can learn from the patterns in the search and application logs about how people are interacting with the application, and with this information, the application can then automatically learn and adjust to those preferences.

Lucidworks Fusion, Signals, and Recommendations

Fusion is Lucidworks' platform for developing enterprise search applications. Fusion is built on three core principles. The first is next-generation relevance, which the leading consumer web search vendors feature today but is still lacking in most enterprise search systems. Second, Fusion is built with the power of best-in-class open source technologies, Apache Solr and Apache Spark. Solr is the open source search engine that has come to dominate the market for search systems similar to how open source Linux has come to dominate the Unix operating system market. Pairing Solr with Spark provides the platform with a large-scale distributed compute engine to apply machine learning and sophisticated processing to traditional search functionality, making possible new features and

capabilities like Signals. Finally, Lucidworks has simplified the development and maintenance of the search system, knowing that most organizations cannot devote a lot of resources to it. Staffing and resource allocation have challenged organizations in the past. Many vendors would claim that their system needed little or no tuning, but administrators find that they have to resort to building rules using these guided navigation systems to achieve the best and most relevant results. In comparison, the use of Signals powered by Spark replaces the manual work needed to build out rule-based systems that were used a decade ago to provide good relevance.

Spark's integration with Fusion's data processing layer enables real-time analytics of application performance and user activity. Lucidworks has integrated Spark into the Solr architecture specifically to accelerate data retrieval and analysis. Developers building with Fusion also have access to Spark's store of machine learning libraries for data-driven analytics.

Fusion provides support for recommendations via the aggregation of signals. Recommendations run against aggregated signals and then are applied to a set of search results from a query. Fusion provides Item-to-Query recommendations (improved query results), Query-to-Item recommendations (top queries that lead to an item), or Item-to-Item recommendations (e.g., "customers who read this also read that").

Lucidworks Fusion's Signals connects to Solr through sophisticated APIs, extending the capabilities of Fusion and making it a powerful platform for a new generation of data-driven applications. Fusion enables teams to quickly build and deploy powerful search applications across the enterprise. In many ways, the enablement of a typical unified information access platform can help with the transition to more targeted, role-specific search-driven applications. Signals provides for more relevant searches as well as recommendations for the "next best action" that users should take. The combination of these capabilities and features provides an extremely powerful system that delivers next-generation relevance and recommendations with less involvement and coding than previous rule-based systems.

CHALLENGES/OPPORTUNITIES

While replacing rule-based systems with signal processing makes a lot of sense and provides a user benefit, it is not without its own costs. The process of activating and configuring Fusion's clickstream capture, processing, and analysis is relatively straightforward. An administrator can simply turn on the recording of user clickstreams, let the data collect and aggregate, and then direct Fusion to apply Signals to analyze the data. If administrators want to go deeper than that, they will have to write their own Spark job or use one of Fusion's existing Spark jobs.

Moving beyond that, to take full advantage of Signals and Recommendations, search administrators need to collect different types of data that can be analyzed and used. Administrators may want to collect information on who collaborates with whom in the office or organization by analyzing email trails or application usage patterns, calendar appointments, and other aspects of collaborative software.

Obviously, this can raise security and privacy concerns. Issues about data security, privacy, and governance will need to be addressed within the organization if tools like Signals are used. Workers in many organizations will need to be assured that this type of data is anonymized and securely held so that individual privacy is not adversely affected. While all of this data is recognized as being owned by the organization, collecting data on email and application usage can still be fraught with problems, outweighing any potential benefit. In many cases, using this tool for public or ecommerce search is much easier as these concerns may not be as prevalent.

CONCLUSION

Search applications are moving from a rule- and heuristic-based methodology for analysis and decision support to an approach where data-based analysis, contextual clues, and machine learning are embedded to provide the next generation of "intelligent" applications. Software vendors are beginning to understand the power and utility of machine learning, signal aggregation, contextual analysis, and other technologies that can create new opportunities for organizations to improve knowledge worker efficiency and value. These technologies are being deployed in a wide number of enterprise applications, especially collaborative and office applications, such as enterprise search. These tools promise to increase the efficiency and value of these software systems while reducing the overhead and upkeep required to maintain them. As IDC has previously noted, unstructured information is 90% of all digital data and is the fastest-growing type of information. Having tools to process, handle, and utilize that information to its fullest extent is crucial for many organizations.

IDC sees a future where almost all enterprise and collaborative applications have elements where the software provides automated assistance, recommendations, and predictions via cognitive or artificial intelligence services. Lucidworks Signals uses these same technologies and capabilities to provide users with sets of recommended answers and next best actions based on their actual usage of the system as well as the wisdom of crowds. The use of machine learning and contextual clue recognition and action greatly improves enterprise search capabilities over rule and traditional relevance approaches that have existed for the past decade. Organizations should strongly consider tools that offer capabilities like these to improve and augment knowledge worker and search user productivity.

Enterprise search is undergoing a fundamental shift from a focus on data to a focus on the user experience. The Lucidworks Fusion platform provides the capabilities needed for teams to build powerful search applications that provide the most relevant results — and a superior experience for the end user.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com