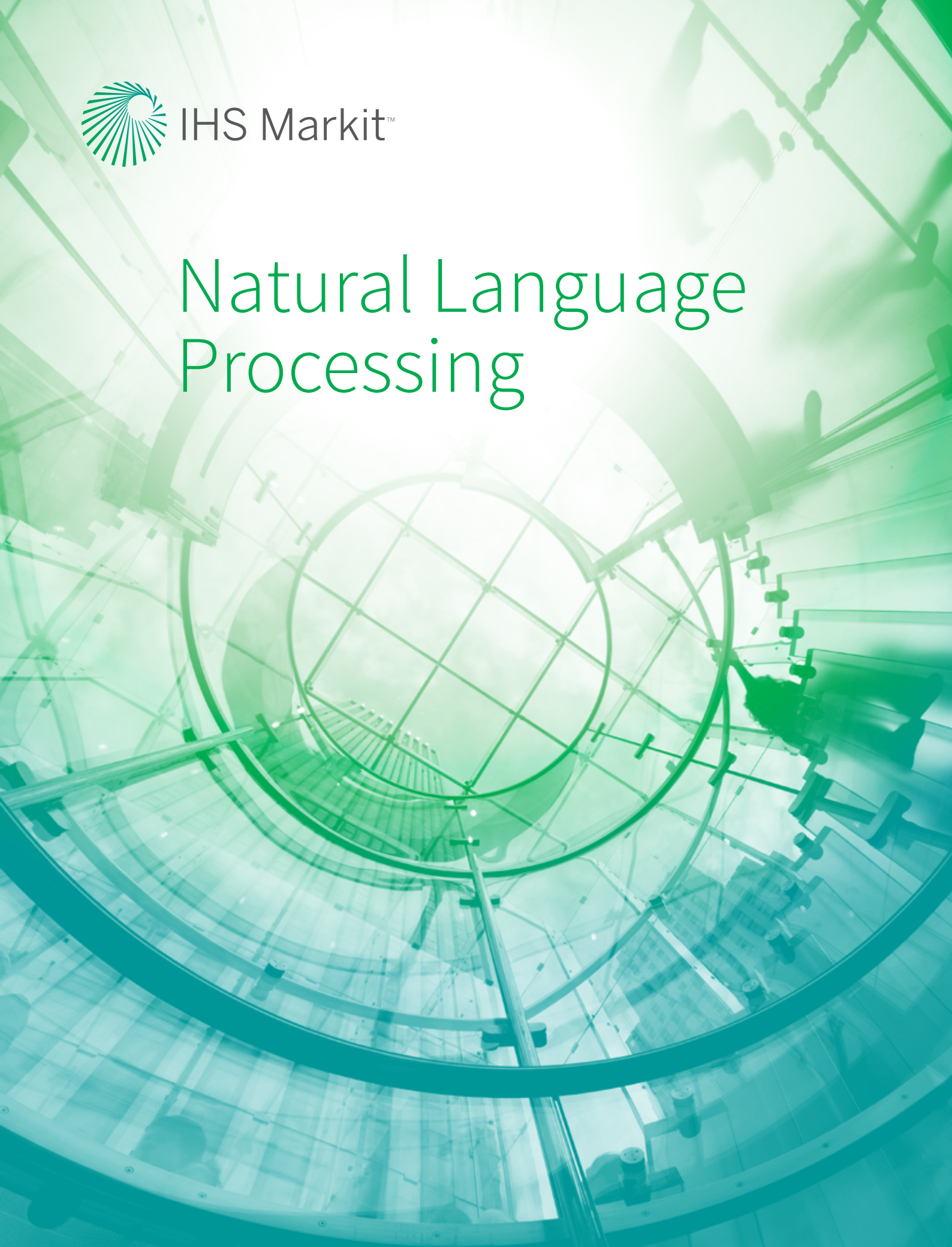
IHS Markit™

Natural Language Processing



Sifting Through Haystacks of Irrelevance: Where can Engineers Turn When Keyword Search Doesn't Work Anymore

How organizations can use advanced knowledge retrieval tools to help engineers, scientists and other technical professionals effectively leverage relevant concepts from both corporate and external content.

Organizations waste up to 30% or more of their engineering and R&D resources duplicating work or repeating past mistakes, according to industry researchers. That's because finding and accessing information stored in repositories scattered across an organization is difficult at best. Yet, this knowledge is critical in an increasingly complex environment, where information from past projects is often buried in decades-old, non-integrated enterprise systems. In fact, analyst firms report that engineers can spend 40% or more of their time searching for information, often searching across a dozen or more systems to find the answer they seek – or worse, failing to find information critical to the project or task at hand.

The fact that data is proliferating at an alarming rate doesn't make the knowledge worker's job any easier. According to IDG, unstructured data (i.e., information that either does not have a pre-defined data model and/or that isn't organized in a pre-defined manner) is growing at the rate of 62% per year, and by 2022, 93% of all data in the digital universe will be unstructured. Gartner's statistics aren't any rosier, with the research firm predicting that data volume is set to grow by 800% over the next five years, with 80% being categorized as unstructured data.

In this article we explore the various ways that engineers are tackling critical research, show where these methods fall short, and explain how advanced knowledge retrieval capabilities help users leverage relevant concepts from corporate and external content.

The Limitations of Keyword Search

In an organizational context, traditional keyword search has proved useful for a narrow set of specific applications namely because it was designed to help knowledge workers find information that they already know exists—a part number, for example, or a named document, or a subject matter expert's name in a document. Both keyword and enterprise search are largely focused on document retrieval and, as such, are extremely limited. With the explosion of information available both inside and outside the organization, these legacy search technologies retrieve piles of documents, as if having the document is enough.

For companies that want to dramatically reduce the amount of time their knowledge workers spend searching for relevant answers, there's semantic search. At its core, a semantic search engine understands the relationships between keywords, phrases, or parts of speech within a search phrase, therefore allowing it to understand the underlying meaning of the entire phrase.

For example, a semantic search engine would be able to easily distinguish the differences between the

following phrases made up of the same “keywords” but with obviously different implications:

- How to burn a dress?
- How to dress a burn?

In this example, the phrases are made up of the same keywords, but the subject-action relationships are reversed. In traditional keyword search—which is based on ranking algorithms—because the relationships between the sentence parts are unknown, a keyword search would return identical or nearly identical results for these two completely different questions.

Additional problems with keyword search arise when the keywords are too specific, producing few or no results, or too general, in which case the results are overwhelming and irrelevant. On the other hand, since semantic search technology understands the meaning of these two sentences, it would produce highly relevant answers to the questions.

The goal of semantics is to always provide the direct insights and answers needed to complete research tasks, rather than burying those ideas among scores of irrelevant documents.

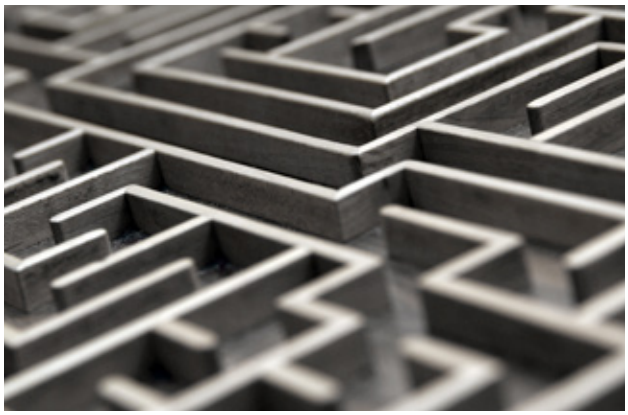
Solving the Knowledge Retrieval Problem

The unstructured text document explosion presents a formidable challenge. At a typical organization, these documents reside within a large number of disparate enterprise systems, folders, and information sources – both inside and outside the organization. As a result, just searching across all these different information sources is problematic.

Furthermore, keyword searches, Boolean searches, and statistically based methods have long been inadequate, retrieving a list of documents that might – just maybe – contain potentially relevant information, and then leaving it to the researcher to pore over each document to determine the value of its contents.

When it comes to innovation and technical problem solving, these conventional search technologies are failed strategies because the process of matching keywords cannot understand the context of the user's request (i.e., their intent and need). That’s because most traditional search technologies produce piles of mostly-irrelevant documents, when what engineers and scientists need are precise answers to specific mission-critical, project-related questions.

Sifting Through Haystacks of Irrelevance



The wealth of corporate wisdom that resides in unstructured documents is often recorded as natural language text. This unstructured content is the motherlode of an organization's' collective tribal wisdom, but mining it with traditional knowledge management search and navigation tools is cumbersome for engineers who need precise insights and who have very limited timeframes in which to retrieve those insights.

The problem is exacerbated when engineers seek out relevant answers from the wealth of information on the web, whether in standards and other regulatory content; within eBook libraries or digital publications; or in worldwide patent collections. Despite the continual evolution of technologies to create and leverage metadata and classification schemes, researchers still find the "needles of wisdom" to be too deeply buried in the "haystacks of irrelevance" to make the effort worthwhile. It's no wonder that research from Infocentric Research shows that 54% of decisions are made with incomplete, inconsistent, inadequate information.

Effective knowledge retrieval demands that computers correctly analyze the user’s information requirements while precisely matching these requirements to the searched documents’ contents. To accomplish this level of intelligence in automation, the source document’s natural language text and the researcher's query must be analyzed into elements that convey meaning. Then, the same methods must be employed as the basis for unambiguous comparison between the query and the documents.

Engineers Need Answers, Not Documents

Unlike keyword search engines that merely catalog a few words, advanced research technology—like the one powering Engineering Workbench by IHS Markit, the leading Engineering Intelligence Platform—indexes and programmatically "reads" the content of documents – understanding and cataloging more than 100 semantic relationships within language.

Developed at a cost of over 1,000 man-years, and with nearly four dozen global patents in the areas of semantic search, question-answering technology, and deep semantics, Engineering Workbench is specifically and purposefully tuned to deliver precise answers to the questions that engineers, scientists, and those in product development ask – and sometimes even the answers to questions they don't know to ask.

Engineering Workbench's natural language query interface helps the user pose questions in free-text format (i.e., the same format as if the question were given to another person). Once relevant knowledge has been retrieved, the software presents the results in a way that makes meanings readily apparent.

By integrating proven problem-solving tools and methods with advanced knowledge retrieval capabilities, Engineering Workbench helps users leverage relevant concepts from corporate and external content. This marriage of high-level concept extraction and problem-solving capabilities ensures engineers, scientists, researchers, and other innovation workers are better informed, more creative and comprehensive in their thinking, and able to make better decisions and solve problems faster.

ihsmarkit.com

About IHS Markit

IHS Markit (Nasdaq: INFO) is a world leader in critical information, analytics and solutions for the major industries and markets that drive economies worldwide. The company delivers next-generation information, analytics and solutions to customers in business, finance and government, improving their operational efficiency and providing deep insights that lead to well-informed, confident decisions. IHS Markit has more than 50,000 key business and government customers, including 85 percent of the Fortune Global 500 and the world's leading financial institutions. Headquartered in London, IHS Markit is committed to sustainable, profitable growth.

CUSTOMER CARE

NORTH AND SOUTH AMERICA
T +1 800 447 2273
+1 303 858 6187 (Outside US/Canada)

EUROPE, MIDDLE EAST AND AFRICA
T +44 1344 328 300

ASIA PACIFIC
T +604 291 3600

E CustomerCare@ihsmarkit.com

174804967-MT-0917