InterSystems

Use Cases for Unstructured Data

Benjamin De Boe
Product Manager
InterSystems Corporation

Use Cases for Unstructured Data

Introduction

Experts estimate that 85% of all data exists in unstructured formats – held in e-mails, documents (contracts, memos, clinical notes, legal briefs), social media feeds, etc. Where structured data typically accounts for quantitative facts, the more interesting and potentially more valuable expert opinions and conclusions are often hidden in these unstructured formats. And with massive volumes of text being generated at unprecedented speed, there's very little chance this information can be made useful without some process of synthesis or automation.

Automating text analysis is not easy. Too often, it requires prior knowledge of what the text is about and a great deal of upfront work to build relevant dictionaries or ontologies. The exception is InterSystems iKnow technology, which employs a unique "bottom-up" approach that analyzes text based solely on what is contained in the text itself.

This paper outlines four basic use cases for the kinds of insights that can be gained from text analysis. For each use case, we briefly describe how real InterSystems customers have leveraged iKnow technology to enhance their applications with text analysis.

The Problem with Top-Down Text Analysis

Text analysis is critical to solving certain kinds of problems, such as:

- ***How can we analyze consumer sentiment for a brand being discussed on a social media platform and track which products or elements it applies to?***
- ***How can population health management in medicine really succeed if the valuable information in clinicians' notes is not incorporated?***
- ***How can documents in a complicated legal matter be correlated to build an effective case?***

In each of these scenarios, the biggest risk with the top-down approach is that information will not be revealed because the process of extraction is biased toward previously chosen words and phrases. Many solutions claiming to handle text indeed only "handle" it by storing and displaying it in full, but lack a true understanding of what's inside: natural language. Search-based solutions typically cling to statistics about previous searches by other users and therefore are inherently biased towards those users' ideas and reasoning. Expert systems combine a machine-interpretable knowledge base with a reasoning engine to draw conclusions from text, but they are limited by the scope of that knowledge base, as well as the budget required to build and maintain it.

There are a few tools that offer true insight into text by starting from the language itself. However, many of these tools are small research or point solutions and interoperability with applications or business processes is often challenging.

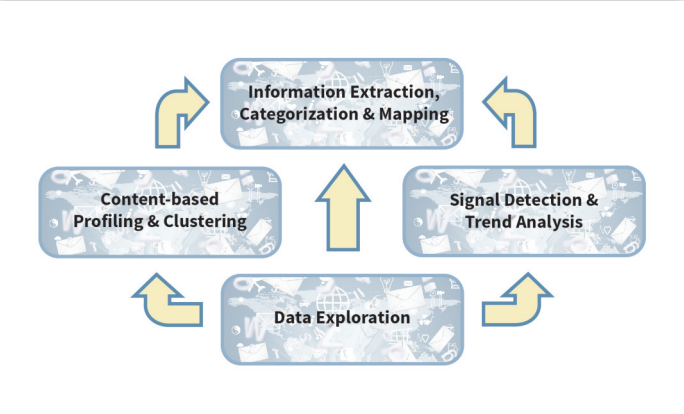
InterSystems iKnow Technology

iKnow is different. Its unique "bottom-up" approach translates free text into quantifiable entities with rich contextual annotations. All of this is solely based on the properties of the natural language in which the text was written and is therefore strictly unbiased.

iKnow technology is embedded into the core of the InterSystems platform, so it is readily available for use by applications and business processes. Whether the context is an enterprise transactional application implemented on InterSystems Caché or a highly-connected, "complete data" environment such as a health information network implemented on InterSystems HealthShare, iKnow provides the ability to analyze unstructured text data.

iKnow Use Cases

There are four basic use cases for text analysis:



In the diagram above, use cases positioned towards the top require more specific domain knowledge and upfront work. Data exploration is typically a purely bottom-up discovery scenario, whereas the other three scenarios require a more specific target to be set by the user. Real-world text analysis solutions probably involve more than one of these use cases, particularly since the results of data exploration are frequently used as input or guidance for the others.

Data Exploration

The idea behind data exploration is simple: it provides insight into possibly large volumes of unstructured data. It is the essential first step in any text analysis project because it helps users evaluate what their data is about and how concepts and topics relate to one another.

Data exploration answers the question:
What are my texts about?

A data exploration solution typically employs some sort of user interface such as a knowledge portal through which users can quickly access the top concepts, see other related concepts (or similar concepts), and identify which documents mention or discuss these concepts. Smart document navigation and rich text search capabilities are often also part of these solutions.

Customer Examples of Data Exploration

Here are two examples of how InterSystems customers are using iKnow for data exploration:

- **AuxiPress**, a media clipping service in Belgium, uses iKnow to track which brand and company names appear in the press. It has grown to be the #1 media analysis choice for political parties in Belgium.
- **CysNET** is a software company in Spain that embeds iKnow in Badakit, a knowledge management solution for healthcare. Badakit enables customers such as the Clinica Universidad de Navarra to build a taxonomy covering all findings or diagnoses in clinicians' actual notes, rather than having to rely on an external database.

Signal Detection and Trend Analysis

A signal detection and trend analysis solution is an advanced version of data exploration used to determine a more precise relevance or otherwise specific attribute of individual documents or a whole body of documents.

A signal detection and trend analysis solution answers the question:
What is truly relevant or specific about my texts?

Rather than relying on frequency metrics, a signal detection solution looks for recurring themes and patterns. It can optionally visualize them in condensed dashboards. Data Exploration is used to explore what the data has to tell by itself, whereas Signal Detection looks for something specific. Examples are sentiment, trigger terms, or other indicators for a particular property associated with each document.

The main advantage of using iKnow for signal detection and trend analysis is that iKnow starts from a richer and more complete representation of free text than is possible with the typical word-based or dictionary-driven approaches. With iKnow, trends are not limited to single words or expressions that were pre-defined as "potentially interesting," and linguistic context such as negation can be appropriately taken into account.

Customer Examples of Signal Detection and Trend Analysis

Here are a two examples of how InterSystems customers are using iKnow for Signal Detection and Trend Analysis:

- **PCS** is a publishing solution provider that embeds iKnow in its Social Knowledge perception management solution. iKnow dynamically identifies themes in social media content about a particular topic, typically an organization, brand, product or service. This helps their brand managers engage directly with the right audience when potential issues or negative themes are detected.
- **Parnassia** is a psychiatry service provider in the Netherlands that uses iKnow to identify trigger terms in clinicians' notes that indicate a high probability of a patient's situation deteriorating to the point of requiring seclusion in a psychiatric facility.

Content-Based Profiling and Clustering

A content-based profiling and clustering solution is used to return documents or groupings of documents that adhere to a certain profile. For example, an online book shop might want to recommend books a reader might also like to read. Or, a mobile news app might want to suggest similar articles for a reader.

A content-based profiling and clustering solution answers the question:
Which texts belong together?

Many content-based profiling and clustering solutions make suggestions and recommendations based on structured metadata, static user preferences, or statistics about other users' behavior. While sometimes effective, these approaches tend to ignore the text itself, which is essentially what the user was looking for in the first place.

The iKnow approach adds structure to unstructured data in such a way that classic algorithms (as well as proprietary InterSystems techniques) can identify similar content or otherwise group documents into sets or clusters. Grouping can be automated, using standard clustering algorithms, or, be fully deterministic, giving users complete control of the criteria that define the sets.

Customer Examples of Content-Based Profiling and Clustering

Here are two examples of how InterSystems customers are using iKnow for content-based profiling and clustering:

- **Koorong Books** is the biggest vendor of religious books in Australia. It uses iKnow to identify duplicate postings of the same books based on their written descriptions, as well as to make reading suggestions of similar books. iKnow is crucial to the suggestion engine because, unlike organizations such as Amazon, Koorong does not have the transaction volume to rely on statistics alone, given the breadth of their portfolio. Nor can it rely on structured metadata in the form of book categories, as it is too labor-intensive and suppliers do not consistently provide the metadata.
- **A number of hospitals and other sites in BeNeLux, Germany, and the US have deployed a "set analysis" application.** This application allows users to define sets within a body of work based on specific terms or combinations of terms appearing in their free text. It is used to identify patients most suitable for clinical trials or other studies from a population, based on a combination of their "structured data properties" and information within the unstructured data associated with each patient.

Information Extraction, Categorization, and Mapping

The fourth use case, information extraction, categorization, and mapping, is about identifying pieces of structured information or metadata within free text. For example, the size of a tumor might be extracted from phrases such as "the tumor diameter was 1cm"; "tumor was estimated to measure between 1cm and 2cm"; or the shorter TNM descriptor found in clinical jargon, "T2N1M3." Other examples of information extraction would be to identify names of locations in text in order to tag them on a map, or categorize surgery reports based on whether the operation was laparoscopic or not.

Information extraction, categorization, and mapping solution answers the question: ***What elements do I recognize in my text?***

In most information extraction, categorization, and mapping solutions, an initial discovery phase using data exploration, and perhaps signal detection are required to build the extraction or categorization model. This model-building phase is independent of running it as part of an application or business process and therefore it can easily be deployed to production environments. Categorization models could also be the end product of a clustering exercise such as with the set analysis tool described in the previous section.

An advanced example of this use case is mapping free text to an established ontology such as UMLS, to identify medications, diagnoses, and other clinically relevant elements. Advanced ontologies such as UMLS also imply hierarchies that add further value when extracting specific information from unstructured text. While these mapping operations can yield multiple elements (for example, multiple medications), it can still be considered a structured metadata property. This does not overlap with the signal detection use case as these mapped elements are returned because they correspond to an entry in the ontology and not merely because of their linguistic role in the text.

Information extraction, categorization, and mapping solutions are usually domain-specific and require some upfront work by the user. However, the language constructs identified by iKnow and the building blocks offered for use by these solutions mimimize the significant effort that would normally be spent creating rigorous and complete sets of regular expressions or vast dictionaries.

Customer Examples of Information Extraction, Categorization, and Mapping

Here are examples of how InterSystems customers are using iKnow for information extraction, categorization, and mapping:

- **PCS** uses iKnow as part of a geo-tagging process in its Knowledge media publishing solution. This enables PCS to assign coordinates to news articles and place them in the appropriate "local news" section of the many titles they publish.
- **N^3** is a software company serving multiple industries. It has implemented information extraction procedures using iKnow for the German real estate site, ImmobilienScout24. The application can derive "missing" structured metadata elements from user-provided textual property descriptions, which are then used by potential buyers for a more focused and effective property search.

Conclusion

Unlike most other text analysis technologies, InterSystems iKnow employs a unique "bottom-up" approach that analyzes text based solely on what is contained in the text itself. It greatly reduces the amount of prior knowledge and effort required to build solutions that leverage information stored in unstructured text fields. Such solutions enable users to make better decisions because they can access and use all their data.

InterSystems iKnow is being used by organizations around the world for data exploration, signal detection and trend analysis, content-based profiling and clustering, and information extraction, categorization, and mapping.

About InterSystems iKnow

InterSystems **iKnow**™ is available as an embedded technology in InterSystems **Caché**® and InterSystems **Ensemble**® enabling any application built on our platform to leverage the value of the unstructured data it manages.

InterSystems iKnow and InterSystems DeepSee power the analytics capabilities of InterSystems' healthcare solutions: InterSystems HealthShare®, a health informatics platform that enables strategic interoperability across a hospital network, community, region or nation; and InterSystems TrakCare®, an Internet-based unified healthcare information system that rapidly delivers the benefits of an electronic patient record.

For more information on how iKnow can unlock text in your application, visit **InterSystems.com**.

About InterSystems

Founded in 1978, InterSystems is a software company with offices in 25 countries and world headquarters in Cambridge, Massachusetts. We develop advanced data management, connectivity, and analytics technologies that help our clients make breakthroughs in healthcare, finance, government, utilities, space exploration, and other industries that demand the highest software performance and reliability. We work directly with leading-edge user organizations as well as innovative solutions vendors that use our technology platform to create breakthrough products and services. InterSystems also provides complete health information systems for hospitals, primary care providers, community care organizations, and laboratories in select countries on five continents.

What sets our company apart is that every employee is dedicated to making clients successful. That is our shared passion, and it drives a continual commitment to excellence in our software products and services. As a consistently profitable privately held company, our focus on client success has never been distracted by demands from outside investors or the short-term gyrations of stock markets.

InterSystems Corporation

World Headquarters
One Memorial Drive
Cambridge, MA 02142-1356
Tel: +1.617.621.0600
Fax: +1.617.494.1631

InterSystems.com