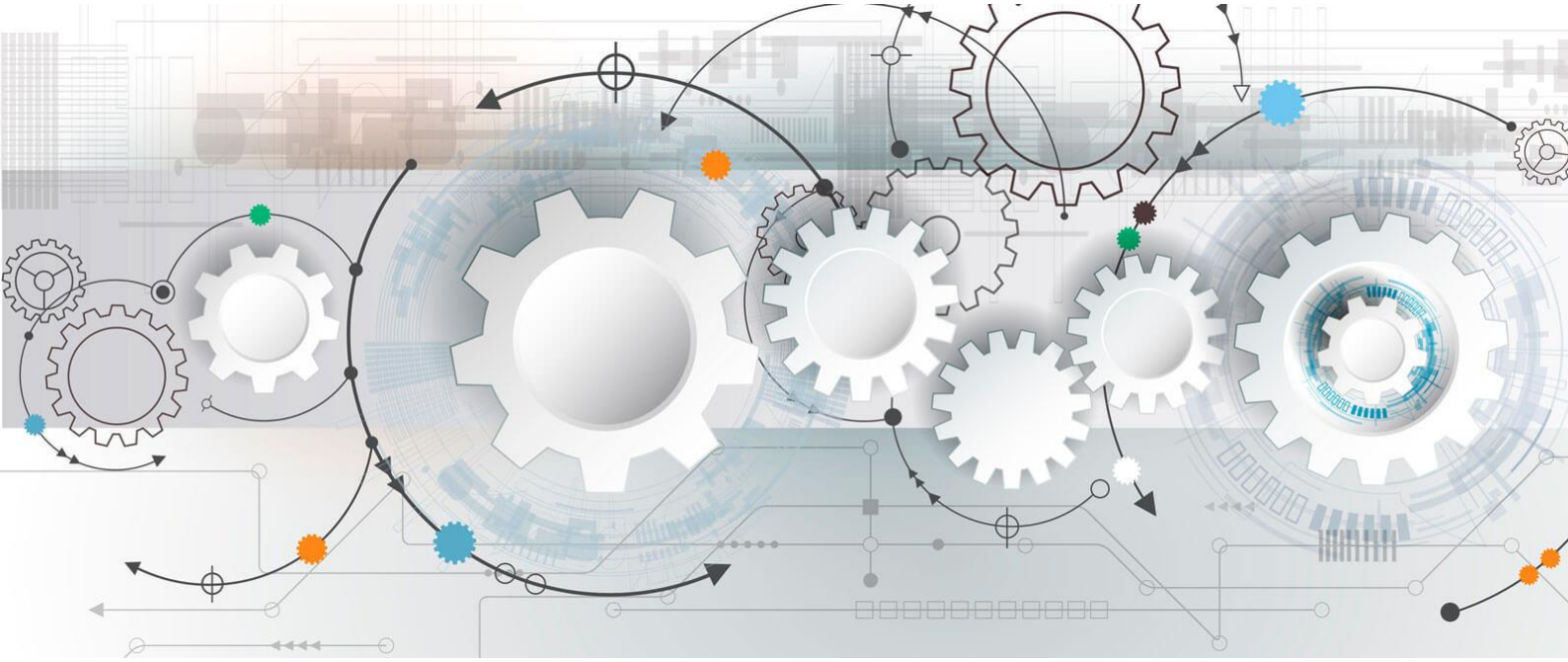
COGITO®: A UNIQUE ARTIFICIAL INTELLIGENCE TECHNOLOGY



Contents

INTRODUCTION

The exponential growth of information and the need to manage its increasing complexity means that traditional technologies alone are no longer sufficient. These technologies, whether keyword or statistic based (such as pure machine learning technologies), are frequently chosen because of their perceived simplicity, but they don't address the depth of today's real-world requirements for an effective search and information management application.

This is true if you consider the traditional areas of application of unstructured information management technologies such as search or knowledge management, but it is especially relevant for other areas in which traditionally unstructured information play a smaller role, such as in strategic decision making.

Decision making once relied on descriptive information models based on ERPs, CRMs or expensive market research and therefore on data collected in the past. The value of such activities was to help understand what might happen in the future based on what already happened. These models are increasingly less effective, not only because they are based on expired data and past observations, but because they are also missing the data stored in unstructured sources—internal, such as communication, collaboration and document management systems including emails, text files, Word documents, presentations, or external, such as information streams, reports and social media —that could turn descriptive models into predictive ones.

These unstructured information sources are an important, if critical part of the equation. The difficulty of leveraging internal knowledge buried in platforms developed and designed just with that purpose, and the strategic importance of news streams and other open information sharing platforms (like social media) have ushered in a greater emphasis for accessing and monitoring these real-time sources of information. To tackle this, innovative technologies that provide a better understanding of concepts and the relationships between concepts, with an emphasis on understanding the meaning of words and all of their nuances cannot be underestimated.

It is precisely here where Artificial Intelligence applied to text shines. In the following pages, you will discover the potential of the Cogito® Artificial Intelligence platform for solving today's biggest information related problems. You will learn:

- How the different legacy technological approaches to the problems of search, classification and automatic data extraction from text documents are increasingly less effective.
- The unique approach of Cogito and its state-of-the-art technological modules.
- The benefits of semantics for making the process of unstructured information management more effective, creating business value for any organization.

KEYWORD-BASED SYSTEMS

Keyword-based systems (also known as full text systems) operate on a "cleaned-up" version of the text, where high-frequency words such as articles and prepositions have been deleted and where a document is filtered as a series of character strings, not words. While this approach offers the advantage of quick indexing, its lack of precision in retrieving search results is a serious disadvantage for advanced search. This leads to both problems of overload—too many irrelevant results—as well as underload—too few results. Its main limitations include the following:

- **Words = Character Strings**: By processing text not as words, but as a series of characters or letters, the meanings of words, their gender, their grammatical and logical role, are ignored. Analysis of a sentence such as "This year we are closing two plants in Italy and opening another two in Poland" would be unable to distinguish between where the plants were closing (Italy) or opening (Poland).

- **No Concept Management**: Without an understanding of meaning, such a system is unable to identify similar concepts present in text. Instead, a keyword search will only retrieve results related to the exact keyword, ignoring related concepts unless they are explicitly stated in the query. To compensate, some systems use synonyms to increase accuracy. This method lessens underload by retrieving more possibly relevant documents, yet it considerably increases overload by finding all of the documents that contain all of the related words, independently from their meaning.

- **Indexing Approach**: This approach organizes an index with the sequences associated with the documents extracted—identifying spaces as separators and including punctuation and other non-alphabetic characters. In search, it indexes the documents returned as a list of search terms, and refers to this index, rather than to any actual documents, in results. In general, it eliminates high frequency keywords present in text because they are considered insignificant and, for the same reason, it does not consider stop words such as articles and prepositions.

- **Ordering of Results**: When searching within a rich set of documents, effectively sorting results is a challenge. Ideally, results should appear ordered by relevancy. Full-text or keyword-based systems use a statistical approach: The greater the number of occurrences of a word in a document, the more importance given to that document in search results. This method is not reliable for every scenario, and requires manual sorting of results.

Keyword technology may be integrated with various tools to overcome these limitations, including ranking algorithms to sort results by relevance, the use of a thesaurus to add synonyms, and the use of stemming to include the root form of words and all related forms *(fishing, fished, fisher, fish)*.

STATISTICS TECHNOLOGY

Statistics technologies refer to all systems based on machine learning, and include systems based on neural networks and fuzzy logic. This approach is based on the statistical inference of the characteristics of a text based on a training set of sample documents or corpus, which is used as a model for how the system should interpret or classify content. Statistics technology may be integrated with other tools to enhance its effectiveness.

PATTERN MATCHING

Pattern matching systems, or those based on neural networks, are based on the same techniques as full-text systems but use more advanced statistical analysis for content, including the following techniques, to improve understanding and text retrieval:

- **Fuzzy Logic**: The use of an intentionally imprecise logic that processes words as a string of characters to account for misspellings or incomplete words. A traditional search engine will return a “Did you mean...?” list in an attempt to guess the right answer. This approach suffers from frequent overload by including words whose root has multiple meanings (For example, the root of pressing (as urgent), is press, which has several different meanings (newspaper, ironing, machine used for printing), and underload in the case of irregular terms where the plural or other tenses are different from the root (knife and knives; go and went).

- **Pattern Matching**: Broadens the word reference unit from a single word to a group of words where the word group is considered a unique object. A collocation such as bird of prey is managed as a unique form and not as individual words. Because the system stores the couple bird/prey as a unique object, it also finds this group in a sentence (“Felines see the birds as prey”) where the meaning is completely different. Because the objects are processed as sequences of characters and not as concepts, birds of prey is not detected as a form of bird of prey, but as a completely different object.

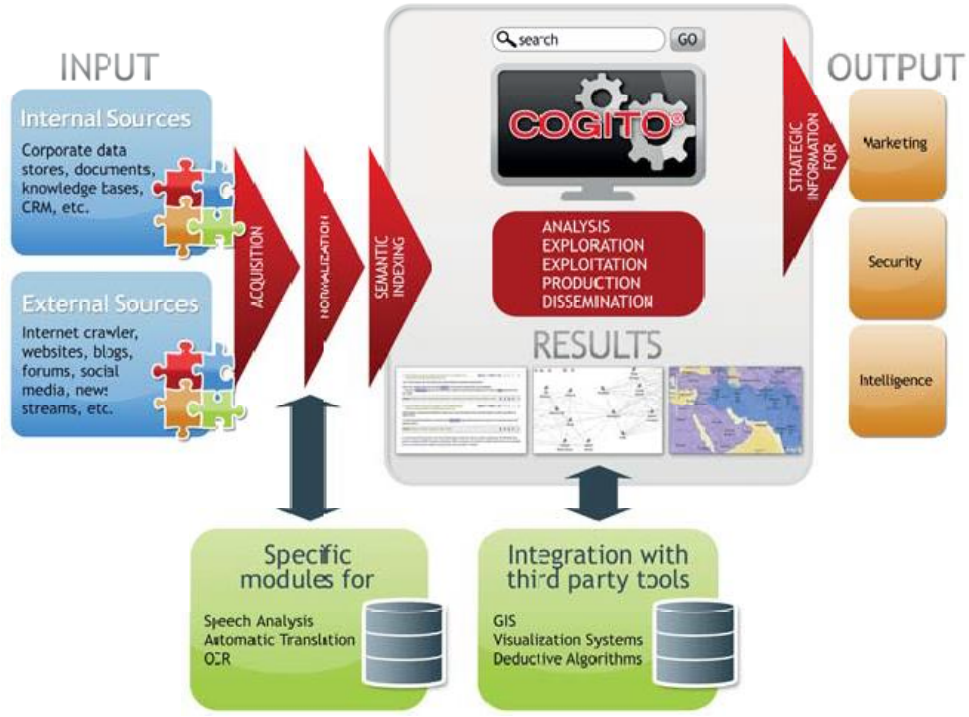
- **Co-occurrence of Terms**: When several words in a group appear together in multiple documents, some level of interdependency is assumed between terms. This method uses a statistical analysis of words that frequently appear together. While this improves classification, it cannot reliably improve search. For example, based on a rule that says documents with score, play and shot are about sports, the system would be unable to disambiguate the different context in the following sentences: (a) The Lakers make a great play and score with a shot from the 3-point line; (b) He poured a shot of whiskey and started to write the score for the play. While co-occurrence of terms may be useful in the search for statistically similar documents, it doesn't enable more accurate searches.

- **Importance of Less Frequent Words**: Keyword systems have a special management for words that are statistically less frequent, operating on the theory that the most informative words of a text are the least common, and consequently must be given more relevance. This can make document classification easier because it exploits a statistically observed feature of texts, but it doesn't necessarily add value to search.

ARTIFICIAL INTELLIGENCE TECHNOLOGY COGITO®

A common characteristic of all the technologies traditionally used to automatically process text is that, acknowledging the limitations of their capabilities, they all choose to create a workaround to approximate the only and most effective way to analyze text, which is to replicate the way people understand text when they read.

The Cogito® cognitive technology can be classified as a software for the understanding and automatic analysis of text. Unlike other technologies, Cogito does not process content as a sequence of characters and does not guess at the meaning of words and concepts. Instead, Cogito is a real AI-based software that relies on a deep semantic analysis and a rich knowledge graph to ensure a complete understanding of a text as a person would.



The result of Cogito's semantic processing is a structured representation of previously unstructured text.

THE KNOWLEDGE GRAPH

All of this is made possible by the Sensigrafo, the knowledge graph at the heart of Cogito. The knowledge graph is a representation of knowledge—the meanings of words, the relationships between concepts—that reflects the richness of how knowledge is structured.

Inside Cogito, knowledge graph lemmas aren't organized in alphabetical order, such as in a dictionary, but in syncons or groups of synonyms that represent the same meaning or concept that the lemmas express. Each syncon is a node in the

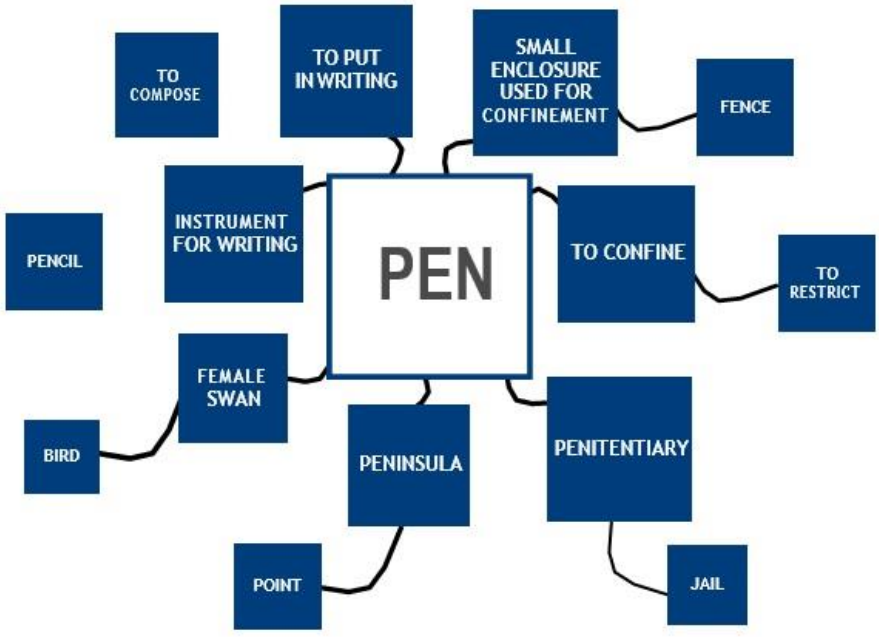
knowledge graph that is linked to other nodes through semantic relationships in a hierarchical structure. In this way, each node, in addition to its meaning and attributes, is enriched by the characteristics and meaning of the nodes that are above or below it. Each of these links identifies a kind of relationship that links the concepts in a language, and the links are principles used to organize the concepts in the knowledge graph. For example, concepts are organized starting from the less specific to the more specific (*vehicle/car/SUV*) or as potential subjects or objects of a verb, etc.

Syncons can contain lemmas, composed words and collocations. The main elements of each syncon are: grammar type, semantic link, the definition/meaning, the domain, and the frequency. In the knowledge graph, the real meaning of each syncon is a combination of its elements (synonyms) and the relationship between the syncons. Cogito's knowledge graph includes different kinds of links between the syncons, which results in a greater ability to represent the understanding of a language.

For a superior understanding of text content, Cogito uses a morphological analysis to process keywords, a grammatical analysis that understands the base lemma, and a logical analysis that identifies the parts of speech in a sentence (subject, verb, object, preposition, etc.)

Cogito's disambiguator analyzes single sentences or entire documents to distinguish the precise meaning for each word from all of the various meanings associated with a term, eliminating any possible ambiguity. In this way, Cogito is able to understand the meaning of words in context to mirror the way humans process information.

The result of Cogito's semantic processing is a cognitive and conceptual map of text—essentially a structured representation of previously unstructured text.



The knowledge graph is a representation of knowledge that includes the meanings of words and the relationships between concepts.

CHARACTERISTICS OF COGITO

Cogito is composed of several integrated elements that are used to disambiguate texts and process natural language, which is essential for the automatic comprehension of a text. Its main components are

- **Parser**: A parser identifies the single elements that constitute a text, assigning them their logical and grammatical value. Consider the examples: (a) There are 40 rows in the table. (b) She rows five times a week. While traditional systems would treat the use of rows equally, Cogito understands their different grammatical role in each sentence, and therefore the different meaning of each. Recognizing a word independently of its written form is important. Cogito distinguishes gender—masculine/feminine—and number—singular /plural—in words and correctly associates all forms of verbs to their common meanings instead of simply identifying them as different words, as other systems do.
- **Knowledge graph**. The core element of Cogito, the knowledge graph, consists of a combination of Lexicon and Knowledge Base, assigning a specific meaning to each word analysed.

- **Lexicon**: Knowledge of all of the possible meanings of words, and in their proper context, is fundamental for processing text content with high precision. Consider all of the various meanings for the word driver to understand how different meanings can make it difficult to correctly interpret content: (a) The driver of the red car was injured in the accident; (b) I used the long driver; (c) The driver was installed in the computer. Representing knowledge in a series of networks, Cogito's knowledge graph represents the complex connections and associations between text essential for disambiguation.

- **Knowledge Base**: Knowledge is a key element for understanding what is being read. A lack of specific knowledge on a subject or area may prevent a person from fully comprehending a text. The knowledge base within Cogito is organized and applied during analysis in a process that can be compared to what humans do when they apply their "common knowledge" to the reading of a text. In the same way our personal knowledge is improved when we learn new things, Cogito's knowledge base may also be increased, enriched and improved by a mechanism of guided learning with current and domain-specific information.

- **Memory**: During document analysis, Cogito uses a retention technique to determine the semantic context of a text, enable disambiguation and consequently extract meaning with the best possible approximation.

CONTENT REPRESENTATION

The outcome of Cogito's analysis is a conceptual map of the text, where:

- Each concept expressed in the text is uniquely identified regardless of which words of a language are used to represent it in the text analyzed.
- Each agent is associated with the action carried out.
- Each object is connected to the related action.

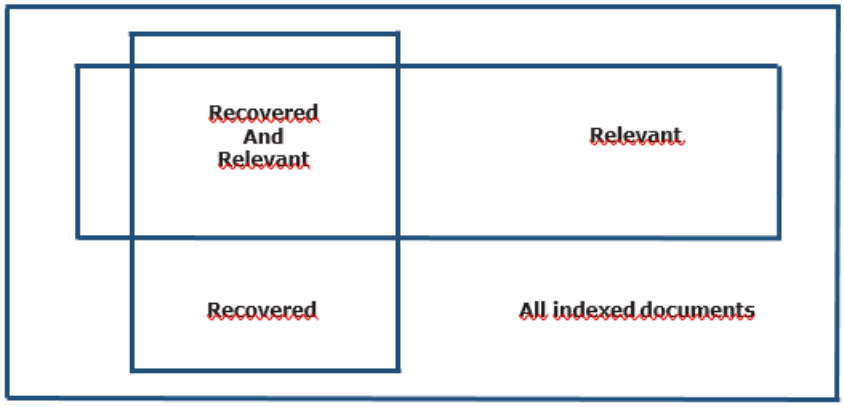
In this representation of content, a document's main topic, as well as other topics, dates, numbers and other meaningful information, may be stored. The map

provides a document with structure, which enables it to be used in formal processing tasks such as indexing, classification, summarization and translation.

TECHNOLOGIES COMPARED

To understand the difference between semantics and other technologies, it is important to compare them in terms of technological effectiveness, as well as from a business point of view.

The effectiveness of information retrieval methods, in other words, relevance, is measured by precision and recall. Precision is the ability to retrieve only the relevant information (for example, in the first page of search results, how many of those results are relevant). Recall is the ability to find as much relevant information as possible (through available domains or repositories) relevant to exactly what you are searching for:



Here, we will look at some of the key advantages that semantics offers for some of the most important functionalities of information retrieval and analysis.

SEARCH

The search function is the software's ability to retrieve documents in a repository such as documents on open sources (such as the internet), as well as those inside an internal business domain. Compared to technologies based on keywords or statistics, Cogito offers some unique features for search. These include:

- **Conceptual search**: A rich knowledge graph retrieves not only documents related to the user's query, but also information that is conceptually similar. Where keyword and statistics technologies employ synonyms and other techniques to approach higher precision and recall in search, they result in a greater number of false results and errors. Cognitive technology enabled by semantics overcomes these challenges because it disambiguates the actual meaning of terms through the power of its knowledge graph. Results are therefore more accurate because they are based on the comprehension of words. For example, in the search for car, Cogito would retrieve not only the word car, but the concept car, which means not only documents that include *automobile*, but also *SUV, sedan,*

convertible, BMW, Mustang, etc., without any manual customization of the index or the query.

- **Natural Language Search**: The conceptual map and triplets—subject, action, object—created as a result of semantic processing enable the understanding of the user's intended query when written in all of the ways that people communicate, including slang, abbreviations, even misspellings. Where a traditional application would process the individual words and ignore the question implied in the following example: *"office hours over the wkend Brooklyn"*, Cogito understands the intended question and returns a set of documents that answer it. Natural Language Search is especially relevant for FAQs, Wiki-like search and search used in customer facing applications online and available via SMS where user communications are often abbreviated).

- **Rich, Customized Ranking Criteria**: The richness of the results derived from semantic analysis of text enables fine tuning of the search platform, as well as implementation of customized ranking mechanisms using flexible and parametric criteria. For example, users in R&D could have the results of a news search for a competitor ranked on how closely the document is related to product innovation while users in Finance could have the same items ranked based on M&A activities or new hires, etc., all information made available in the analysis phase.

CATEGORIZATION

As companies around the world strive to find better options for managing the inflow of unstructured information, they often turn to classification systems such as taxonomies to organize the chaos by recognizing and differentiating content. Automatic categorization of text according to a predefined taxonomy is a common method for efficiently organizing information for retrieval.

The development of classification systems and the management of data has quickly become a science. Generally speaking, a classification system will contain several parts: 1) The collection itself, 2) A classification hierarchy (tree) that categorizes documents by topic, 3) Sample documents describing the type of content to be classified within each category/node of the hierarchy and 4) An information platform that drives collection of content from the appropriate data sources and then places the content in the correct category.

As with other typical application of technologies for natural language processing, categorization also has three main approaches:

- **Manual**: Also known as the "bag of keywords" approach, this requires compiling a list of key terms that describes the type of content in question, and uses them to develop a metadata list. If a piece of content is assigned to a node, it contains the keywords assigned to that node.

- **Statistical:** Statistical systems use a training set of documents that talks about the same topic and uses different algorithms (Bayesian, LSA or many others) to extract sets of key elements of the document to build implicit rules that will then be used to classify content.

- **Semantic-Based:** Linguistic rules are written to capture the elements of a document that assign it to a category; rules can be written manually or generated with an automatic analysis and only then validated manually (a

time savings of up to 90%). Rules are deterministic (you know why a document is assigned to a category), flexible and powerful for an effective categorization. An approach based on semantics offers an important advantage for categorization. Performance is not dependent on the size of the taxonomy because the knowledge present in the knowledge graph (each concept includes many attributes and links to other concepts that can be leveraged in rule writing) enables a better understanding of the rules in context and the ability to more easily refine the rules.

This means that once the system is deployed, documents that do not "fit" into a specific category are identified and automatically separated, and the rule developer is able to fully understand why it was not classified. The administrator can then make an informed decision about whether to modify an existing rule for the future, or create a new class for content that was previously unidentified.

Other technologies, especially those that are statistics-based, reach a plateau at a certain level of taxonomy complexity (usually up to 200 categories). For larger taxonomies, the cost of managing it becomes too high or the precision becomes too low to be used in real-world information management scenarios.
Specifically:

- Because content in the information being gathered is ever changing, keyword systems must be continuously updated to stay relevant. This is the oldest method but also the most time consuming and inefficient, not to mention unscalable.

- Statistics-based systems have no understanding of words, nor can the system administrator pinpoint why specific terms are selected by the algorithm or why they are being weighted. In the event that the classification is incorrect, there is no accurate way to modify a rule for better results. The only option is to select new training documents and start again. New rules must be constantly recreated to keep up with changes in the content being ingested.

The cost of implementing a taxonomy of categories with semantic technology is, on average and given the same level of performance, about two times lower than other technologies. One reason is its ability for leveraging the knowledge in the knowledge graph that makes it faster to reach the required objective.

Semantics can address the ambiguities in language, and is able to scale the process of identifying and organizing a list of preferred and alternate terms, which is a labor-intensive process with other approaches.

While a statistics approach may seem fully automatic because training documents can be used to add to the taxonomy, it ignores two important drawbacks: 1) Most organizations do not have a training set of documents for each node of the category, which inevitably causes accuracy and scalability problems; 2) Training actually requires a significant amount of time and manual work. Instead, semantics does not have these limitations.

TEXT MINING

Text Mining extracts specific information from text, normalizes it in a standard format, and archives it in a database. Semantic technologies excel in text mining by providing:

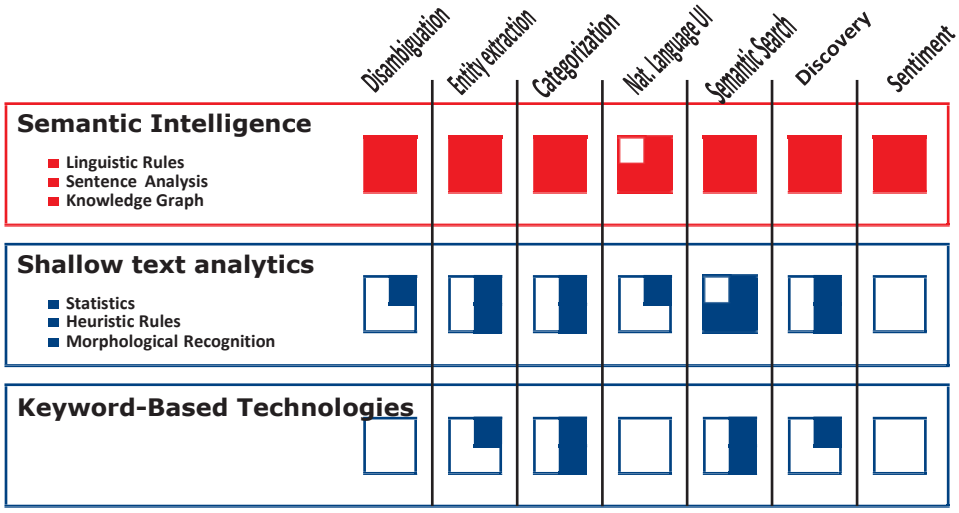
- **Higher Precision and Recall in Entity Extraction**: Semantics ensures higher precision and recall in text mining functions thanks to its out-of-the-box recognition of tens of thousands of concepts enabled by the rich knowledge graph. By leveraging the conceptual map generated by the semantic analysis of text, semantics is able to provide the additional contextual information to validate the identification of information to be extracted. In addition, extraction performance does not depend on whether entities are known or unknown ahead of processing and does not require writing a line of code or adding a single term to the lexicon.

- **Shorter Time and Lower Costs**: Because of its ability to automatically perform text mining, shorter implementation times and therefore lower implementation costs are significant advantages offered by semantics when implementing a custom extraction project.

- **Triples Identification and Export, and Compliance with W3C Standards**: The entities and concepts extracted by the system can be linked with one or more of the tens of thousands of relations automatically recognized during analysis and exported in RDF format.

- **More Precise, Relevant and Rich Tagging**: The elements of the text are ranked based on semantic relevance and on the role played in the text. This ensures, together with the other steps of deep understanding, a more effective automatic tagging of the analyzed content.

- **Easy Implementation of Custom Ontologies**: The richness and completeness of the core knowledge graph and compliance with W3C standards enables a minimum amount of customization work when compared with more basic technologies.



CONCLUSION

The increasing quantity of information that must be processed and understood on a daily basis, combined with the need to lower the costs of computing power and storage are just a few of the challenges facing organizations. Traditional technologies that were once favored due to speed and easy application are unable to meet the information management needs that today's complex and ever-growing information sources require.

Semantic technologies offer the advantage of a deep, contextual understanding of information without compromising quality in terms of precision and recall of information processed. Instead, they ensure an automatic comprehension of content that is significantly higher than alternative technologies. Not only is semantics the best solution in the typical activities of information retrieval, categorization and extraction, it also offers the most promising approach in the development of more advanced applications (both enterprise and consumer) and integration with the most advanced predictive models.

www.expertsystem.com
info@expertsystem.com