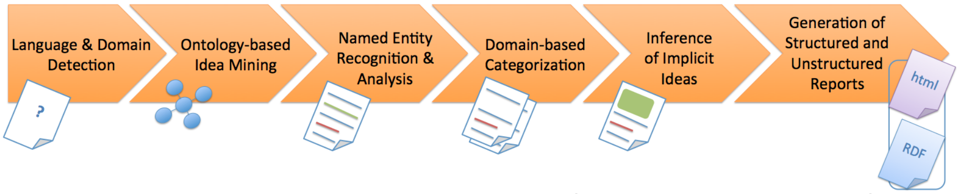
IntuScan™ DOCEX

Document Exploitation

IntuScan™ DOCEX is an integrated semantics-driven platform for real-time exploitation of unstructured textual documents, which integrates intuition and knowledge of seasoned experts into an automated process.

The purpose of IntuScan™ DOCEX is to extract all relevant information from a large quantity of unstructured texts written in a variety of languages, and to generate a structured representation and natural language report, which includes the characterization of the document, identified entities, and other information implicit in the document.

IntuView
Pioneering artificial intuition


Language & Domain Detection
Ontology-based Idea Mining
Named Entity Recognition & Analysis
Domain-based Categorization
Inference of Implicit Ideas
Generation of Structured and Unstructured Reports
?
html
RDF

Language and Domain Detection

IntuScan™ DOCEX begins the analysis process by identifying the languages of a given document sorted by their prominence. IntuScan™ currently identifies more than 60 languages; some of them may use the same script. IntuScan™ also distinguishes between parts of the same texts written in different languages, and identifies "loan words" and phrases in "hybrid languages" – concepts from a "guest" language that are transliterated and integrated into the "host" languages – and restores them to the source language for extraction of meaning. For some of the languages, IntuScan™ DOCEX identifies the general domain of the document (e.g. drugs-trafficking, chemicals, terror, etc.) allowing the following phases to use a fine-grained logic that matches the specific domain.

Idea Mining

IntuScan™ DOCEX analyzes the text using its powerful integrated Natural Language Processing (NLP) engine to uncover language independent concepts that are defined in an ontology for various domains. As opposed to existing approaches, which follow the "bag-of-words" idea, IntuScan™ DOCEX uses sophisticated linguistic tools to identify complex expressions and to resolve ambiguous expressions according to their domain context. IntuView has developed in-depth ontologies for several domains relating to the fields of homeland-security, finance, and law-enforcement. Unlike traditional flat ontology methods, the IntuScan™ ontology is a multi-dimensional structure of relationships between unique concepts. Users can work with the provided ontologies or extend and enrich them to meet their specific requirements, by using a knowledge management tool.

Named Entity Recognition and Analysis

IntuScan™ DOCEX recognizes all the named entities (e.g. persons, organizations, locations, facilities, dates, events) existing in the document, transliterates them into various standards, and analyzes them to find the roles of each entity name part. The recognition component is based on a hybrid approach that combines statistical models along with rules, based on linguistic and cultural knowledge (naming conventions, etc.) for identifying and classifying entities. This method extracts implicit information derived from the names (gender, ethnicity, etc.) and contextual information surrounding them (titles,

IntuScan™ DOCEX is available for the following verticals:

- Homeland Security
- Law Enforcement
- Finance

IntuScan™ DOCEX currently supports the following languages:

Arabic, English, French, Indonesian, Spanish, and Urdu

IntuView is an innovative company dedicated to the development of "artificial intuition" software for security and defense applications.

The IntuView vision is to **revolutionize knowledge mining** and **cross-language** extraction of information by replacing current technology of language-dependent, generic lexical searches with language-independent, domain-oriented "idea mining".
The new technology will provide the knowledge consumers with summaries in their own language of information gleaned from a broad spectrum of sources in different languages.

types, sentiment, etc.) in order to aggregate and disambiguate entities and to find affinities between them. The extracted information is modeled according to the ontology and formatted as RDF structures. The information can then be used by IntuScan™ Name Matcher to match the extracted entities with existing lists.

Domain-based Categorization

IntuScan™ DOCEX categorizes a given document according to a predefined list of categories and values. The categories are carefully selected for each domain as part of the domain ontology. The values are assigned based on statistical calculations that aim to find the similarity of the given document to a corpus of manually annotated documents of the same category. The similarity is measured using the ideas, entities, and other information extracted from the document in previous steps.

Inference of Implicit Ideas

IntuScan™ DOCEX uses the extracted entities, ideas, categories, and the general context in which they occur to discover additional implicit information such as sentiment, affiliation, and key themes of the analyzed document. This is performed by ontology-based rules that capture domain-specific patterns and structures. The rules can be easily modified to comply with the user's specific requirements.

Generation of Structured and Unstructured Reports

IntuScan™ DOCEX generates a natural language report for any given document. The report characterizes the document and presents the key ideas in their appropriate context. The report is currently available in English and French. This information is also generated in a RDF structure that is stored in a triple-store semantic database for future query. Currently IntuScan™ DOCEX is fully integrated with Oracle 11*g* and other database platforms.

IntuScan™ DOCEX currently supports the following languages: Arabic, English, French, Indonesian, Spanish, and Urdu. The following languages will be available in the near future: German, Persian, Russian, Somali, and Turkish.

IntuScan™

For more information or to request an evaluation copy, please write to info@intuview.com