

Semantex™

Enabling Information Discovery

What if you could efficiently maximize data to be more useful, relevant and timely? Janya's Semantex™ automatically does this, and more. Approximately 70 percent of data encountered every day comes from unstructured sources. Semantex effectively exploits information fully, utilizing the most sophisticated techniques available to extract all relevant meaning, identify relationships and put them in a form that can be acted on in a timely manner. It doesn't simply catalog facts, it provides an evidence trail and connections among relevant information. The information extracted can support other discovery activities including enhanced enterprise search, advanced keyword search with text analytics, link analysis and timeline visualization. Most important, Janya's Semantex creates a more complete picture from your data, so you can work effectively.

Product Overview

Semantex is a complete information extraction system that supports automatic or semi-automatic analysis of large volumes of electronic information, detecting entities, attributes, relationships and events. It uses a hybrid model for information extraction, combining machine learning with natural language processing to maximize strengths and achieve the greatest accuracy. Semantex is modular, scalable and portable to many domains, and has been used to power both commercial and intelligence applications. Using a distributed architecture, Semantex scales easily with multiple processors to achieve the throughput necessary for high-volume, data-intensive government applications. The XML output reflects rich features assigned to elements of text based on syntactic and semantic knowledge.

Entity Profiles

Entity profiles represent the main output of Semantex. Unlike other systems that are limited to local entity extraction, Semantex consolidates all the information about an entity into a profile, generating a profile for each unique entity in the document. Each entity profile contains all the extracted information about a specific **entity**, including its **attributes**, **relationships** with other entities and **events** involving the entity. By combining all the information for a given entity in a profile, Semantex™ provides access to the extracted data for more advanced investigation, such as hyperlink browsing and chase-the-rabbit scenarios. Merged profiles span multiple documents and maintain original document connections.

Entities

Semantex supports the automatic identification and tagging of entities, including named entities ("Osama bin Laden") and nominal or unnamed entities ("the terrorist"). Native support is provided for over 50 entity types and subtypes in several classes, including person, organization and location. In addition to basic name recognition, Semantex can detect alternate forms of entities, and treat any alias as an additional mention of the entity. Semantex also performs coreference resolution, i.e. identifying pronouns and nominal references and relating them back to named entities, e.g. "he" associated with "John Smith". Alias and coreference resolution together ensure more complete entity extraction and accurate entity counts, eliminating multiple named and unnamed references returned as different entities. Cross-document entity disambiguation helps insure that merged profiles all refer to the same entity, regardless of source document.

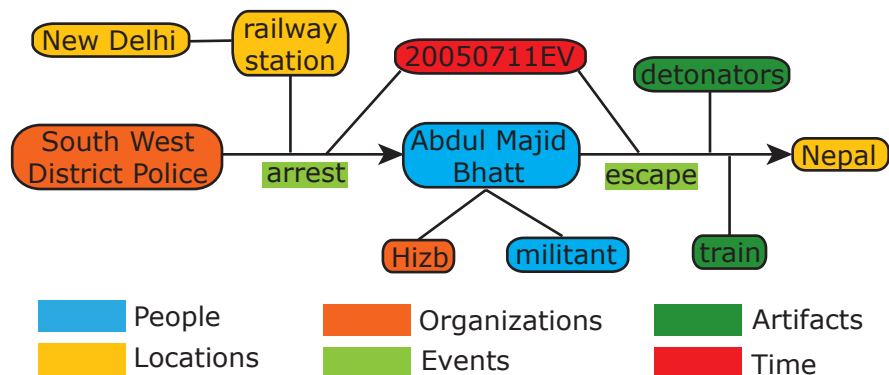
Relationships and Attributes

Complete and accurate detection of entities is only the beginning. Semantex™ detects the key attributes of entities and relationships between entities that contribute to the complete picture of the information. Native

Product Features

- Automatic extraction of:
 - Entities
 - Attributes
 - Relationships
 - Events
- Cross-Document Entity Profiles
- Date/Time normalization in TIMEX2 format
- Scalable, fault-tolerant architecture
- Supports text, HTML and XML input, including NewsML
- Support for input from a variety of external sources, including:
 - USMTF
 - HUMINT
 - FBIS
 - Factive, LexisNexis
- Out-of-the-box ontology based on intelligence domain requirements

"Hizb militant Abdul Majid Bhatt was arrested Monday evening by South West District Police from the New Delhi railway station with three detonators as he was about to escape by train to Nepal."



Sample Semantex™ Output

support is included for a wide variety of business and inter-personal relationships, with over 50 predefined relationships included out of the box. These relationships can be viewed from different vertical domains (intelligence, business) or categories of relationships (business, family, criminal). Also included are quotes and general descriptive phrases that relate back to entities. Entity detection with the addition of relationships and attributes gives a comprehensive summary of each extracted entity found in a document.

Events

Information extraction is incomplete without the ability to discover the events

that associate entities together in real world activity. Semantex™ can perform general event detection based on key verbs, identifying the generic roles associated with the verb: the "who", "what", "where" and "when". In addition, Semantex™ can also detect and extract domain specific events, with native support for more than 40 events in business and intelligence domains. Semantex also detects full Subject-Verb-Object-Complement (SVOC) relationships to add further detail and clarity to the extracted information. Together with relationships, events give users at-a-glance views of all salient activity the entity has been involved in, and can help to constrain search or filter other discovery tasks.

Date/Time Normalization

Making sense of widely varying time references in documents can be a challenging task. Semantex simplifies that task by performing date and time normalization for any references identified, assigning absolute date and/or time information for relative entries and standardizing the format for all explicit date and time references to the TIMEX2 standard.

Architecture

Ever changing needs require a system that can adapt as your requirements change. Semantex was designed from the beginning with a flexible, scalable and fault-tolerant architecture. The core system uses a distributed architecture that can accommodate multiple document processing modules, either on a single server or multiple servers. Increasing throughput is simply a matter of adding another processing module. The intelligent control system adapts immediately to the addition of a new processor, or the loss of an existing processor due to hardware failure, redistributing the work to other available processors. The processing modules use a multi-level approach, combining a variety of techniques – machine learning, rule-based and lexicon/grammar-based – to most efficiently and completely extract information from your documents. This processing pipeline is very flexible, and can easily be adapted to new domains or analysis needs, and can support multiple domains simultaneously.

Technical Specifications

- Dual Processor CPU, 2.0 GHz or better
- 1 GB RAM minimum, 2 GB recommended
- 300 MB Hard Disk Space
- Linux RedHat EL3 & EL4
- Windows XP Pro
- Solaris 8 & 9
- MySQL Database Support

Flexible Customization

While Semantex provides a wide range of native support for entities, relationships and events, Janya recognizes that every domain is different. With the available Semantex Workbench™ tools, you can customize the processing to fit a new domain or modify an existing one. If you don't have the time or the expertise to perform the customization, Janya has a staff of highly qualified specialists that can assist you with domain customization.