ICE – Intelligent Content Engineering

Klaus Netter, DNC Dr. Netter Consulting GmbH, Saarbrücken
Hannes Meyer, RC AG, Fassberg-Müden

ICE (Intelligent Content Engineering) is an industrial strength platform for the automatic disclosure and analysis of textual content with methods from Language Technology and Artificial Intelligence. The platform is based on UIMA (Unstructured Information Management Architecture) and integrates a broad range of different analysis functions, which can be combined and configured in different workflows.

The application can be modularly extended and comprises the following functions:

- Fully automatic text classification on the basis of rules or self-learning AI-components
- Identification of topic-clusters or topic sequences
- Configurable recognition of (near-) duplicates
- Extraction of named entities and information extraction (under development)
- Language identification and other pre-processing modules

For the realization of these functions different components are available which are either proprietary or licensed from other providers. Each of the functions integrated into ICE can be configured and parameterized through a web based administration user interface.

Specialized tools are available for the development and management of models applied in the processing, such as category trees, sets of training or test data, hierarchical rule sets, etc. A comprehensive version management supports the controlled development of models. There is a strict separation between the development and the deployment of models, i.e. the application of the data in a processing environment. Thus, different versions of a model can be used at the same time and independent from each other for different modes (e.g. a testing and operational mode) as well as on different instances of the system distributed across the network.

In addition, ICE offers a comprehensive environment for the testing, benchmarking or evaluation of the individual analysis functions. This allows the user to inspect processing results at all levels, to track the effects of changes and to evaluate the performance in general. The operative workflow as well as individual test runs can be analyzed either at document level or at the level of sets of documents through detail or overview statistics.

ICE and UIMA

For the development of ICE, the UIMA basis played an important role. UIMA simplifies the development of analysis components and allows unifying them to a large extent. On the basis of the principle of chained components, the so-called Aggregate Analysis Engines, complex tasks, such as categorization, segmentation, tokenization, language recognition, etc., can be distributed over several logical components. This yields self contained and reusable components.

ICE benefits from the strength of UIMA's configuration management to respond as fast as possible to client's requirements and to easily extend ICE and its analysis processes. The lightweight approach makes it easy to integrate UIMA into different deployment scenarios and technologies, such as EJB, Spring etc.

This approach also helps during the development phase of new Analysis Engines, since testing UIMA components is done easily and fast inside a Java IDE – no special container or runtime is needed. Additionally, UIMA's Eclipse Plug-in helps visualizing and creating Analysis Engines, Type Systems etc.

The separation of code and configuration of a certain Annotator allows to change the configuration during application runtime – a so called “deployment”. The Annotator will automatically notice changes in configuration and reload itself or the whole dependent Analysis Engine. This approach also allows adding or removing whole components from Analysis Engines – e.g. remove Document Consumers for Statistics, Indexation etc.

Use Cases

ICE is an essential component inside a distributed document workflow system called **CEP** – Content Enrichment Platform – which incorporates (among others) components for the acquisition and normalization of data sources, the central storage and administration of documents, the content analysis, workflow controlling , as well as the delivery interface to different ‘consumers’ such as retrieval, CMS or DMS systems.

Embedded into the CEP, ICE is currently being used by a large German chemical company for a worldwide news management system. This news system continuously collects documents from over 2000 content sources, which are analyzed and classified according to a hierarchical category model. The users can choose from these categories those subjects which are relevant from them. The corresponding newsletters on the basis of 40.000 personal profiles are then distributed on a daily basis to the employees across three different time zones. One of the biggest challenges in this case was the controlled migration, restructuring and redefinition of the category model in ICE without affecting the existing search profiles.

ICE will also be employed by a multilingual public radio and TV station for the fully automatic recognition of topic chains in newswire reports from three different languages. The objective is both to cluster the news items and to filter them for long term archiving. The task here was not only to automatically relate the sometimes rather heterogeneous texts across different news agencies, but also to account for several complex systems of filter and selection rules defining negative or positive representatives within the clusters for the long term archiving. Practically all of the necessary rules and specifications could be defined and configured in ICE’s modeling components in a declarative way.

Another application example for ICE (and CEP) is a web monitoring system combined with a CMS-System, which systematically surveys online sources, checks for relevant content and classifies the retrieved texts for further distribution.