Expert System

Automatic classification of documents for the Library of the Chamber of Deputies

Bruno Bassi
Senior Consultant
bbassi@expertsystem.it
+39 333 9130697
The Library of the Chamber of Deputies, which is responsible for managing and collecting the legislative documents and parliamentary records from Italy and many other countries, needed to support its existing manual method of document classification with a classification system that is compliant with EuroVoc standards.

EuroVoc is the multilingual, multidisciplinary thesaurus covering the activities of the European Union, the European Parliament in particular. It contains terms in 22 EU languages, and is used by national and regional parliaments in Europe, as well as national governments and private users around the world.
The solution

• In 2008, the Library of the Chamber of Deputies introduced a solution for the automatic classification of documents based on a semantic approach, with the objective of optimizing the archival process and conforming to the EuroVoc standards on a European level (composed of 6645 categories).

• They selected the COGITO Semantic Technology solution from Expert System to support the cataloguing and classification of these documents.

• Today, using COGITO, each document is analyzed and interpreted in order to be quickly archived in the corresponding category.
Atto Camera

Interrogazione a risposta orale 3-00799
presentata da
ANTONIO MEREU
giovedì 26 novembre 2009, seduta n.253

MEREU. - Al Ministro dell'ambiente e della tutela del territorio e del mare. - Per sapere - premesso che:

con la delibera del Consiglio dei ministri del 30 novembre 1990 veniva istituita l'area ad elevato rischio di crisi ambientale del Sulcis-Iglesiente, comprendente i comuni di Portoscuro, Gonnosa, Carbonia, S. Antioco e San Giovanni Suergiu;

con il decreto del Presidente del Consiglio dei ministri del 23 aprile 1993 è stato approvato il piano di disinquinamento, al fine di attivare un sistema di risanamento territoriale, con particolare riferimento alle fasce circostanti l'area di Portoscuro, Portovesme e Gonnosa, a seguito dei risultati sui livelli di contaminazione da metalli pesanti effettuati dall'azienda sanitaria su tali territori;

la giunta regionale sarda, con la deliberazione n. 6/56 del 2003, ha dichiarato il territorio del Sulcis-Iglesiente «Area ad elevato rischio di crisi ambientale» con validità quinquennale, che risulta attualmente in esecuzione.

Destinatari:
Ministero destinatario:

• MINISTERO DELL'AMBIENTE E DELLA TUTELA DEL TERRITORIO E DEL MARE

Stato iter: IN CORSO

Classificazione EUROVOC:

CONCETTUALE:
IMPAATTO AMBIENTALE, INQUINAMENTO INDUSTRIALE, INQUINAMENTO RADIOATTIVO, MALATTIA, METALLO PESANTE, RISANAMENTO, RISCHIO SANITARIO, SVILUPPO ECONOMICO, SVILUPPO SOSTENIBILE, TASSO DI INQUINAMENTO

SIGLA O DENOMINAZIONE: SULCIS
What is Semantic Technology

There are three generations in the development of software solution for processing the knowledge expressed within texts:

- Keyword technology or statistics
- Shallow linguistic technology
- Semantic Technology
The limits of traditional approaches

**Keyword technology or statistics**

Breaks text into single words without considering the context—like reading a language that we don’t understand:

*Az IBM szokásosan nagy hangsúlyt helyez a továbbképzésre, így munkatársai évente számos szakmai tanfolyamon vesznek részt.*

**Shallow linguistic technology**

Recognizes words and brings them back to their most basic form (lemma), but cannot distinguish between different meanings.

*Sell -> selling -> sold*

Neither understands the *meaning* of words.
Why semantics is different

- Understands the relationships between words
  *Luke (subject) has eaten (verb) a chicken (object).*

- Understands the meaning of words
  *“To eat” in the sense of to ingest (the chicken); to consume (oil); to destroy (sweater); to run through (money); to rust (the tower)...*
What is a semantic network?

The heart of semantic technology:
- Quality of results derived from the complexity and richness of the network.
- Includes all definitions of all words.
- Include relationships among all words.

**COGITO® English Standard Semantic Network:**
- 350,000 words
- 2.8m relationships
Disambiguation demo

1.
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

2.
Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust,
The automatic classification process

The process of automatic classification works as follows:

• **Incoming Text** requires classification

• **Linguistic Analysis and Disambiguation** provides both grammatical and logical analysis, and associates the text elements to individual concepts (i.e. determines the correct meaning for each word)

• **Linguistic Rules** are then applied in order to achieve categorization: every rule defines a series of conditions that are matched against the actual meaning of the text.

• **Results of classification**: each document is “tagged” with one or more categories (primary or secondary) based on its meaning.
Beyond automatic classification, the semantic approach has also been proven effective and affordable in other automatic tasks, such as:

- Extracting relations between items in texts
- Checking the compliance of documents with respect to a given standard
- Automatically linking related documents
- Perform search based on semantic criteria
- Enabling natural language search (handling “spontaneous” user input)
Thanks for your attention!

Bruno Bassi
bbassi@expertsystem.it
+39 333 9130697