CyMON - SDMS

Semantic Document Management System



Why SDMS ?

The information society is built upon the availability and utilization of the huge amount of information. Especially with the penetration of the world-wide-web infrastructure and the associated technology, efficient access to and the deployment of the correct information play nowadays a key role in the success or failure of most business activities.

Information Overload

The major part of the information is presented and managed as textual documents in numerous languages, formats and styles. The *abundance of information* in the information society has brought, besides the enormous chance for successful businesses, also the new challenge for managing such documents. Among others, it means the complexity and the higher demand on resources and expertise for retrieving the right information from the world-wide database of documents.

The Relevance Issue

According to some market analysis, employees of many companies typically spend up to 80% (depending on the nature of the businesses) of their working time on searching for *relevant and useful information*. Errors, delays and frustration resulted from lack of human resources or expertise are typically the key issues in realizing efficient and successful business workflows.

Personalization

The shift from the provider-centric to the consumer-centric business models requires for the *support of multiple views* of the same documents, according to the individual needs, preferences and expertise of each user or user-category. As such, automation of the generation of taxonomies and advanced semantic personalization tools become highly relevant business resources and enablers for a knowledge intensive organisation.

To cope up with this new challenge, sophisticated and intelligent automatic tools for assisting or replacing humans will become an indispensable part of any document management workflows.

Deployment of statistical technologies for machine learning and data processing contributes significantly to the current success of most commercial document management systems. However, to achieve the performance and quality levels required in a variety of application contexts, extensive utilization of the linguistic semantic knowledge encoded in the documents becomes a key approach.

SDMS focuses on integrating statistical and semantic Natural Language Understanding / Natural Language Processing (NLU / NLP) technologies in realizing a precise, efficient and flexible document management system for a wide range of application areas.

Anatomy of SDMS

SDMS has a three-tier architecture: the SDMS kernel tier, the SDMS shell tier and the SDMS applications tier.

The **SDMS kernel** realizes the basic functions for document management, using the semantic linguistic knowledge and the statistical machine learning/data processing technologies. Three groups of such functions are supported:

- **document classification**, which enables the automatic association of existing documents to some predefined hierarchical/ontological categories.
- **information clustering**, which focuses on grouping the information or documents into clusters based on the semantic similarity of their contents. Such grouping can be used, among others, to support the dynamic creation of ontology/categories taxonomies for documents or document collections.
- **information extraction/summarization**, which enables the automatic extraction of key information from the documents, and enables the efficient assessment of document content at a conceptual level.

The **SDMS shell** provides the interfaces for configuring, combining and utilizing the basic management functions in different contexts. The graphical user interface and the Java API play an important role in integrating and deploying SDMS in heterogeneous environments. SDMS can be deployed as a *stand alone* or as a *server application* that can be accessed via a number of popular protocols (HTTP, RMI, etc.)



SDMS applications can be implemented on top of the SDMS shell. Some typical applications in this context are:

- **document retrieval**, where the result of clustering and classifications can be used to organize the document information using the ontological concepts, and allows the users to browse/retrieve relevant information using semantic concepts. Classification and clustering can also be further used to create models of the documents in a specific category. Such a model can be used for an extended query to search for information from a broader database, e.g. the web.

- **profile learning**, where the categorisation result is used to assess the content categories of the information the user frequently retrieves (see Agentscape Cyb/CyMON agent technology), and used in this way to learn the user's preference profile. This function can be extended to the production & publishing activities and used for learning the users' expertise profiles.

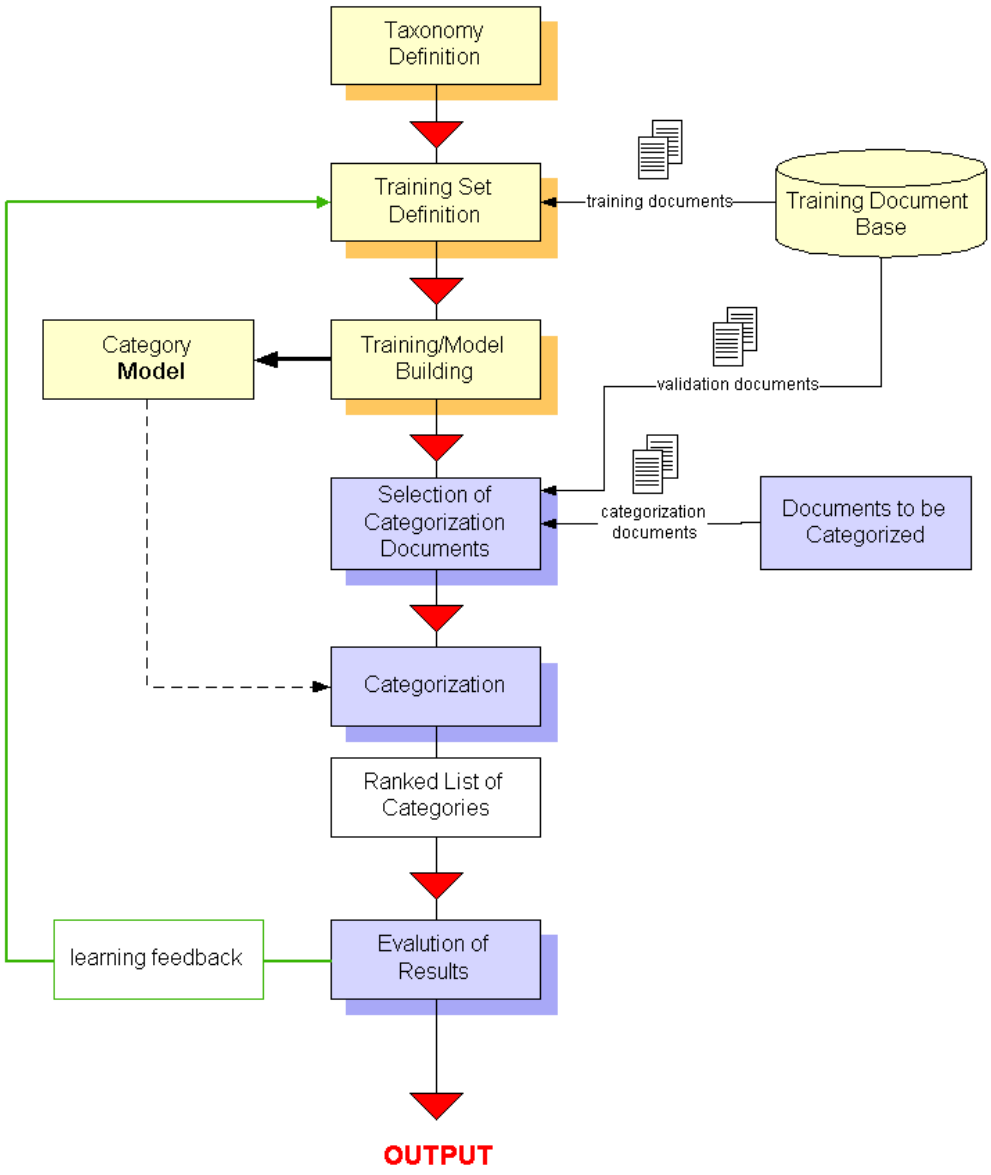
- **response generation,** where the results of classification and information extraction can be used to support some automatic pre-processing of documents in order to decide on some initial responses. E.g. the burden of call centres can be significantly alleviated if a software agent can be in the position to generate some (initial/preliminary) responses to the users and/or to route the incoming message to appropriate destinations.

Categorizer Module

The Categorizer can be viewed as a *knowledge definition and indexing tool*. The information the categorization model can built may be composed of various types of knowledge.

The categorization process starts with a dynamically defined taxonomy of categories and builds up a model for this taxonomy – training process.

Following the model building phase new documents can be categorized - run-time process, with a typical performance of more than 100.000 documents per hour.



main workflow for the SDMS

Statistical processing

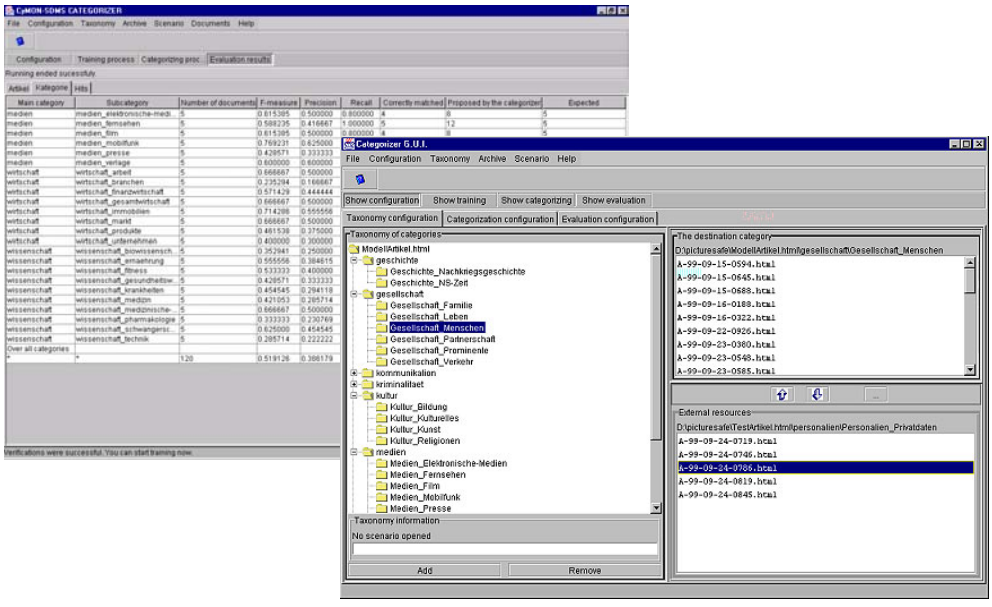
Basically, the SDMS Categorizer is organised around statistical measures it retrieves from the documents.

Linguistic processing

In case of highly flexional languages such as German a pure statistical approach will fail since the concepts need to be mapped to the same definition no matter what flexion they might have. Linguistic processing comes to be essential to such languages. Tests on German datasets with the SDMS Categorizer showed that our linguistic processing raised the quality of categorization with more than 20 percentages.

Semantic processing

At categorization run-time, an optional module uploads semantic knowledge resources in form of ontologies (e.g. synonym, hyponym, or domain-specific). These type of ontologies help the categorizer to map a set of concepts that actually never appear in the training process (model building) to some concept that was defined during model building. This semantic level of processing succeeds in making categorization decisions more accurate.



The semantic processing of the basic unit of written or spoken text actually represents the next step in categorization products. Depending on what is considered to be the basic unit of text (concept, phrase, paragraph) SDMS Categorizer can make the step towards a much more refined semantic knowledge processing.

Who Should Use SDMS ?

The following customer groups will most benefit from the current SDMS development:

Enterprises looking for an intelligent and flexible solution for corporate knowledge management, where the SDMS can be deployed, e.g. as a server, to categorize, organize and maintain textual documents, and to distribute the relevant information in appropriate forms (e.g. in abstract) to the appropriate employees according to the enterprise policies.

Portal providers which can use the SDMS for automatically publishing and organising the web pages and for supporting intelligent browsing and searches.

Call and Contact Centres where SDMS can be deployed to categorize the incoming messages and to forward such messages to the proper / responsible operators. Furthermore the SDMS can support the automatic generation of initial responses based on the semantic knowledge extracted, on the profile / history of the user and on an available Domain Knowledge Base (e.g. FAQs).

Xtreme Personalisation

How can you contact us?

Address

Agentscape AG
Bülowstraße 66
10783 Berlin, Germany

Fon +49-30-59 00 478-0
Fax +49-30-59 00 478-99

Contact

http://www.agentscape.de
info@agentscape.de