Russian Academy of Sciences
Program Systems Institute

Text Mining: Knowledge Extraction and Management

Overview

Why text mining?

Information is overwhelming

Information is crucial to decision-making

Only part of relevant information can be found in the form of *structured data* (databases, spreadsheets, Web forms etc.)

Unstructured data, or texts, remain the premier source of vital information

Problem Statement

Large amounts of unstructured text content in various fields

- E-Science
- Business Intelligence

Natural language processing (NLP) is computationally intensive

Data mining approach applies to structured content only

Goals

Combine different NLP techniques to provide an efficient knowledge management solution

- text categorization
- information extraction
- information retrieval

Use GRID technology for distributed natural language text processing and storage

Enhance existing text mining technology

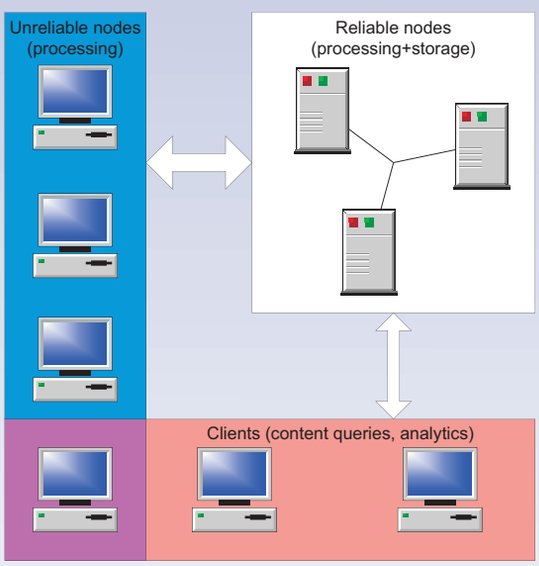
Issues

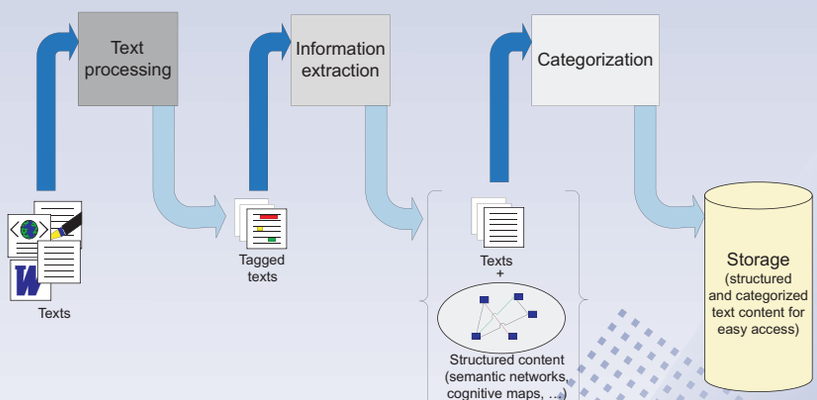
- Portability between domains (e.g. documents in e-Science and Business Intelligence domains differ a lot in language, concepts and writing style)
- Handling multilingual content
- Poor quality of real-life texts (documents may include speech transcripts, optically recognized texts etc.)
- Software portability

Status

Technologies and software tools developed:

- INEX a portable system for information extraction
- AKTIS a portable text categorization tool
- TGRID distributed computations in a cross-platform network environment





ADDRESS

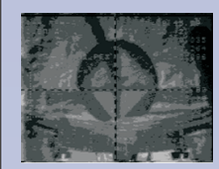
Artificial Intelligence
Research Center
Program Systems Institute
Russian Academy of Sciences

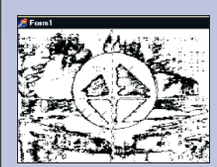
Pereslavl-Zalessky
Yaroslavl Region
Russia, 152020
Tel/Fax: **+7 (48535) 98065**
E-mail: airec@botik.ru
Web-site: Http://www.botik.ru/PSI/AIReC

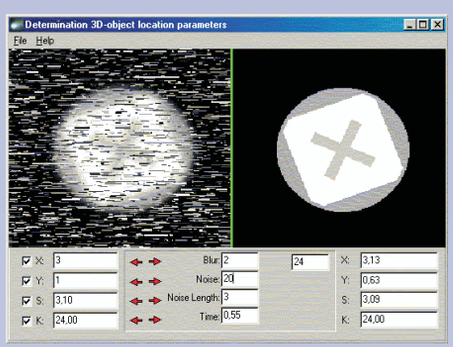
Russian Academy of Sciences
Program Systems Institute

Software System for Docking Target Parameter Identification

Developed under the contract with the Russian Rocket and Space Corporation “Energia”, this software system is meant to identify parameters of the target construction of docking unit directly from TV or video image.





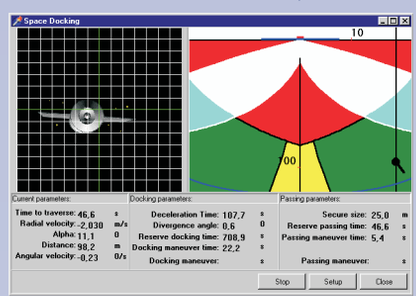


Major advantages of the system are:

- low sensitivity to image blur and noise
- the system is operational even if image is more than 50% truncated

Intelligent Control System for Spacecraft Docking

The software system developed in cooperation with the Russia's Gagarin Cosmonaut Training Centre is meant to ensure safe spacecraft docking under manual piloting. The relative motion of spacecraft is represented by hodographs on an error plane.

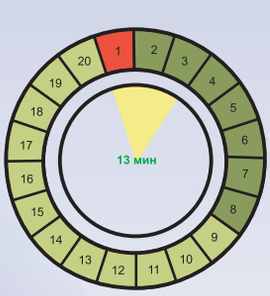


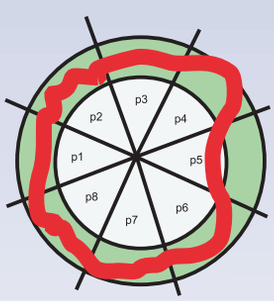
Key features of the control system are:

- informative and intuitive visualization based on hodographs
- real-time situation analysis and forecast
- possibility of using telemetry data to review and analyze real-life docking procedures

System for Visualized Analysis and Control of the Launch Preparation Process

The software was developed under the contract with the Rocket and Space Corporation “Energia”. The objective is to dynamically visualize processes involved in spacecraft launch preparation. An off-nominal process is marked in red and detailed by a red curve where it goes outside the bounding circles.



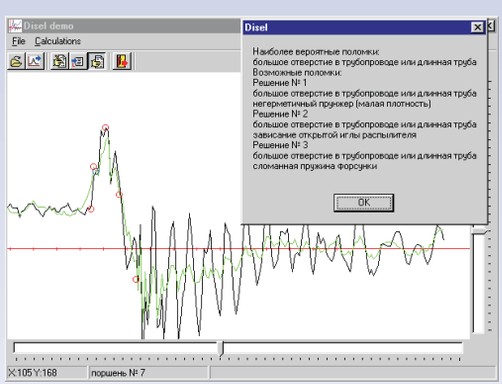


Advantages of the system include:

- efficient, in terms of perception time, visualization of information
- considerably decreasing psychological load on human operator
- ensuring highly informative visualization

Automated Diagnostics System for Diesel Engine Fuel Equipment

The software developed for “Lokomotiv” corporation uses SIMER+MIR expert system to perform fuel equipment diagnostics for diesel engines.



Major benefits:

- prompt failure diagnostics and localization
- significantly reducing equipment down-time

ADDRESS

Artificial Intelligence
Research Center
Program Systems Institute
Russian Academy of Sciences

Pereslavl-Zalessky
Yaroslavl Region
Russia, 152020
Tel/Fax: **+7 (48535) 98065**
E-mail: airec@botik.ru
Web-site: Http://www.botik.ru/PSI/AIReC