An Approach to Context-driven Document Analysis and Understanding

DAS 2000

Claudia Wenzel, Heiko Maus

German Research Center for Artificial Intelligence (DFKI)
Department of Document Analysis & Information Management
Context-driven DAU within the VirtualOffice project

- Motivation
- System overview
- Implementation of the context usage
- Conclusion
Goals to be achieved in VirtualOffice

- Enhancement of DAU by considering background knowledge, i.e., context
- Seamless integration of DAU-system into Workflow Management Systems (WfMS)
- Assignment of incoming documents to corresponding workflow instances
Context serves as information source for DAU

CONTEXT CONCEPT

expectations:
• contents and meaning of an expected document
• consists of contents data and reference data

information need:
• list of needed data items
• DAU derives analysis task
VirtualOffice allows for a flexible analysis of documents in business processes

SYSTEM OVERVIEW

DAU control

plans

resources

analysis specialists

document knowledge & analysis data

corporate databases

WFMS

expectations

DAU components
Context is delivered from workflow instances

- Workflow instances
  - Add context pools
  - Context pool
    - Inference engine
      - Facts
        - Domain ontologies
        - Transformation rules
          - Context unit
            - Analysis task
            - DAU control
              - Document knowledge
                - Contents data
                - Reference data
                - Administrative data
                - Information need
  - DAU
    - Generate expectations
      - DFKI
Uniform representation formalism for generic document knowledge, database schemas, expectations, and analysis results

DOCUMENT KNOWLEDGE

- **Document Knowledge**
  - **Paper Format**
  - **Document**
  - **Paper Letter**
  - **Invoice**
  - **Invoice of GM**
  - **Text-Block**
  - **Sender**
  - **Record Data**
  - **Date**
  - **Clerk**
  - **Clerk: Herzog, Zimmer**
  - **Herzog**
  - **Supplier Database**

- **Relationships**
  - **Is-a**
  - **Has-Dataface**
  - **Has-Parts**
  - **Instance-of**

- **Concepts**
  - **Concept**
  - **Dataface**
  - **Expectation**
  - **Instance**
Conclusion

The system architecture for context-driven DAU results in:

• faster system runs  
  (e.g., reducing search space, extracting only requested data)
• higher precision (e.g., by using background knowledge)
• flexible analysis tasks (given by the workflow)

Moreover, the system

• is domain and WfMS-vendor independent  
  (e.g., generic document knowledge; context pool)
• integrates paper-based documents into workflow instances
TREATMENT OF EXPECTATIONS

expectations restrict the range of possible values within concepts

1) range restrictions after analysis
2) range restrictions while collecting the information from document knowledge
3) do 1), but use more generic concepts as well and give lower certainties to instances of concepts
4) ignore expectations
   (simulates the “ordinary” way of document analysis)
Dynamic control

ANALYSIS CONTROL

Plan-1:Plan( name="ProcessDetermination", next={ start-1 })
start-1:Start( next={ spezialist-1 })
ende-1:Ende( prev={ spezialist-16 })
...
spezialist-8:Spezialist( name="SenderExtraction", next={ entscheider-1 },
prev={ spezialist-6 })
entscheider-1:Entscheider( wenn="Sender unique?", dann={ spezialist-2 },
sonst={ spezialist-9 }, next={ spezialist-2, spezialist-9 })
spezialist-9:Spezialist( name="SenderEvaluation", next={ entscheider-2 },
prev={ entscheider-1 })
entscheider-2:Entscheider( wenn="Sender rating satisfying?", dann={ spezialist-11 },
sonst={ spezialist-10 }, next={ spezialist-10, spezialist-11 })
spezialist-10:Spezialist( name="ShortSenderExtraction", next={ spezialist-11 },
prev={ entscheider-2 })
...
spezialist-13:Spezialist( name="ProductDataExtraction", next={ spezialist-14 },
prev={ spezialist-12 })
...
feedbackChannel = -k  LEFT-SLOT = layout.left
orbHost = "-x"  TOP-SLOT = "layout.top"
orbPort = "-y"  RIGHT-SLOT = "layout.right"
remote = "serv-401"  BOTTOM-SLOT = "layout.bottom"
...
CONCEPT-PARTS = ( "Papier-Geschäftsbrief.Vorgangsangaben.Kundennummer"
"Papier-Geschäftsbrief.Vorgangsangaben.Belegvorgangsnummer"
"Papier-Geschäftsbrief.Vorgangsangaben.Bezugsvorgangsnummer" )
DAU components and resulting specialists

- Logo recognition
- Color image segmentation
- OCR, voting, and lexical processing
- Text classification
- Information extraction by pattern matching
- Knowledge-based table analysis
- Company logo
- Product class
- Message type
- Record data
- Reference data
- Product data
- Process data
- Sender
- Recipient
- Product tables
attributes represented:
slots for:
• image (color)
• layout (coordinates, page)
• content (thesaurus relations, language)
• logic (patterns, grammars)
• message (domain-specific information units)
• dataface (database columns)
facets for:
• uncertainty
• inheritance
• linkage to thesauri
• values, types
Results

PROCESS DETERMINATION

test with 12 documents

without expectations:
  • wrong process determination in 4 cases

with expectations:
  • 18 expectations from 6 different processes
  • process determination always correct
  • shorter runtime of specialists (e.g., logo recognition, pattern matcher)
Results

PATTERN MATCHER

Investigation of reasons for errors within 250 documents:

„conventional“ concept-based analysis:
error rates : 17-43%

use of expectations:
error rates decrease from 20-45%.

The reason for that is that a lot of errors depended on character recognition problems for single characters (which can be nearly totally repaired by an error-tolerant matching based on expectations).