Implementation of a distributed test environment for medical datasources

David Goeth

University of Passau
Graduated seminar: Distributed Information Systems
Supervisor: Armelle Natacha Ndjafa Yakou
WS 2017/2018

February 12, 2018

Overview

Motivation

- Vast amount of data in the fields of medicine
- Data is heterogeneous (e.g. using different file formats)
- Data isn't easily accessible and needs lots of tools
- Research and health care institutions would benefit from a unique view of data [RR14]:
  - Lower Costs
  - Detecting diseases at early stages
  - Simplified collaboration
  - Health care fraud detection

Data Integration

- Data integration addresses the integration of data sources into one information system [LN06]
- Unified view of the data
- Uses a global schema local data source schemas have to be translated to.
- Types of integration: Materialized, Virtual, Hybrid

Heterogeneous Data Sources

- Data sources providing not the same methods, models and structures for accessing their data
- Technical: Query language, exchange format, communication protocol,...
- Syntactic: Number formats, character encoding, tab vs. comma separated values in CSV files $\Rightarrow$ In general: Technical differences in the presentation of information
- Data model: relational, object-oriented, graph-based,...
- Semantic: Differences in the interpretation/meaning of the data
- Structural: Schemas are different though they are semantically equal

Ontology

- Defines the concepts, relationships, and other distinctions being relevant for modeling a domain [LL09].
- Operates on the semantic level rather than on the logical/physical one.
- Independent from lower data models
- Can be used to solve semantic heterogeneity: Logical Inference
- Suitable for integrating heterogeneous data sources

Dataspace

- New abstraction of data management [FHM05]
- Data to be managed rarely stored using a single data model
- Supports all kind of data rather than only a few
- Provides tools allowing a tighter data integration process when required
- Difference to a Data Integration System (DIS): Data coexistence. DIS needs a semantic integration process before providing any services.
- Pay-as-you-go integration
- Multimedia dataspace model[NKCB11]
    - classes, objects and relations
    - Similarity relations
    - Dataspace views
    - Meets requirements for multimedia data

Project Overview
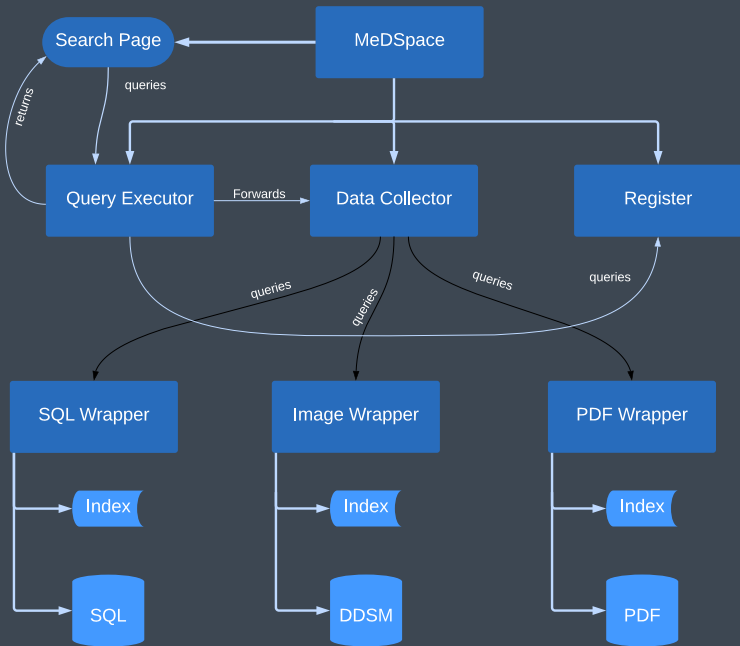
Implementation
Project goals

- Setting up three heterogeneous data sources containing medical data about breast cancer (Relational, Images, PDF Files)
- Implementing wrappers adding keyword search functionality to the data sources
- Wrappers register themselves to the Register
- Convert data of search results to RDF
- All data sources can be queried through an unified user interface

RDF

- Resource Description Framework (RDF) [rdf14]
- Graph-based abstract data model
- Facilitates merging data expressed in different schemas
- RDF4J (formerly Sesame) framework used in the thesis' project [RDF18]

RDF
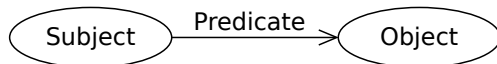
Figure: Connection between two nodes (Subject, Object) forming an RDF Triple [2]

- Three kinds of nodes: IRI, literals, blank nodes
- Relations between nodes are specified by predicates
- Predicates are IRIs themselves
- Web Ontology Language (OWL) can be used to define ontologies using rdf [owl12]

---

[2] https://www.w3.org/TR/rdf11-concepts/rdf-graph.svg

Wrapper (general)

- Proxy for a datasource
- Transforms an incoming dataspace query into a query the wrapped datasource understands.
- Offers its services through REST services, powered by the Play framework [Pla]
- Registers and deregisters autonomously to the dataspace
- Has to offer at least a keyword search query service
- Apache Lucene used to offer fulltext search functionality [Luc16]
- Keyword search currently only supports OR boolean operator
- On startup index for the keyword search service is created

SQL Wrapper

- Sample data used from Schmidbauer's PDGF [Sch12]
- sql data is translated into rdf equivalent using Bizer's D2R Map language [Biz03]
- relational to rdf mappings are defined in seperate configuration file

Image Wrapper

- Digital Database for Screening Mammographyc(DDSM)[DDS]
- Mammography Images used to seek for breast cancer
- DDSM-Utility used to convert (old) JPEG1 files to PNG [Sha16]

PDF Wrapper

- Apache PDFBox [PDF18] for extracting unstructured text content from pdf files.
- The pdf files were formerly created using Schmidbauer's medical sql data set [Sch12]

MeDSpace

- Responsible for managing registered datasources
- Provides user interface for executing search queries
- User Interface and REST services powered by Play framework [Pla]
- Keyword search queries are forwarded to each datasource. The result is collected by the data-collector module
- Query results are (temporarily) stored in separate RDF repositories

Improvements and Outlook

- Query caching
- Incremental (Pay-as-you-go) data integration
- Inferring knowledge using ontologies
- Query language for multimedia data [DYT$^+$09]

References I

Christian Bizer, *D2r map a database to rdf mapping language*, 01 2003.

University of south florida digital mammography home page, http://marathon.csee.usf.edu/Mammography/Database.html, Last accessed on 02/10/2018.

M. Doller, A. N. N. Yakou, R. Tous, J. Delgado, M. Gruhne, M. Choi, and T. B. Lim, *Semantic mpeg query format validation and processing*, IEEE MultiMedia **16** (2009), no. 4, 22–33.

Bhakti Mehta Ed Ort, *Java architecture for xml binding (jaxb)*, http://www.oracle.com/technetwork/articles/javase/index-140168.html, March 2003, Last accessed on 02/10/2018.

References II

📄 Michael Franklin, Alon Halevy, and David Maier, *From databases to dataspaces: A new abstraction for information management*, SIGMOD Rec. **34** (2005), no. 4, 27–33.

📄 M. Tamer zsu Ling Liu (ed.), *Encyclopedia of database systems*, 1 ed., Springer, 2009.

📄 Ulf Leser and Felix Naumann, *Informationsintegration: Architekturen und methoden zur integration verteilter und heterogener datenquellen*, dpunkt, 2006.

📄 *Apache lucene core*, https://lucene.apache.org/core/, 2016, Last accessed on 02/10/2018.

📄 A. Ndjafa, Harald Kosch, D. Coquil, and L. Brunie, *Towards a model for multimedia dataspaces*, Multimedia on the Web (MMWeb), 2011 Workshop on, Sept 2011, pp. 33–37.

References III

📄 *Owl 2 web ontology language document overview (second edition)*, https://www.w3.org/TR/owl2-overview/, December 2012, Last accessed on 02/10/2018.

📄 *Apache pdfbox*, https://pdfbox.apache.org/, 2018, Last accessed on 02/10/2018.

📄 *Play framework*, https://www.playframework.com/, Last accessed on 02/10/2018.

📄 *Rdf 1.1 concepts and abstract syntax*, https://www.w3.org/TR/rdf11-concepts/, February 2014, Last accessed on 02/10/2018.

📄 *The eclipse rdf4j framework*, http://rdf4j.org/, 2018, Last accessed on 02/10/2018.

References IV

📄 Wullianallur Raghupathi and Viju Raghupathi, *Big data analytics in healthcare: promise and potential*, Health Information Science and Systems **2** (2014), no. 1, 3.

📄 Maximilian Schmidbauer, *Medical data generator*, Unpublished Bachelor thesis; Accessible through the libary of the university of passau, 2012.

📄 Anmol Sharma, *University of south florida digital mammography home page*, https://github.com/trane293/DDSMUtility, 2016, Last accessed on 02/10/2018.