Hochschule Zittau/Görlitz
UNIVERSITY OF APPLIED SCIENCES

Enterprise Application Development Group http://www.hszg.de/ead
Faculty of Electrical Engineering and Computer Science

Networked XML Topic Maps (NXTM)

A project for extracting structured data from unstructured data.

Jörg Lässig, Adam Bartusiak, Florian Haje

University of Applied Sciences Zittau/Görlitz, Department of Computer Science, Görlitz, Germany

Introduction

- common information overload is a significant problem nowadays
- in the IT universe 80-90% of digital data is unstructured
- unstructued data is rather intended for human consumption only:
  - it has no pre-defined data model
  - it is not organised in a pre-defined manner
- the usage of existing data search tools for unstructured data is limited:
  - it is difficult to discover, collect and extract valuable information

Goals

- extraction of structured data from unstructured data from multiple resources:
  - emails and text messages
  - MS Office and PDF documents
  - XML and HTML files
- dynamic recognition and representation of linked information in documents
- flexible and intuitive graphical user interface enabling easy access to the analyzed data

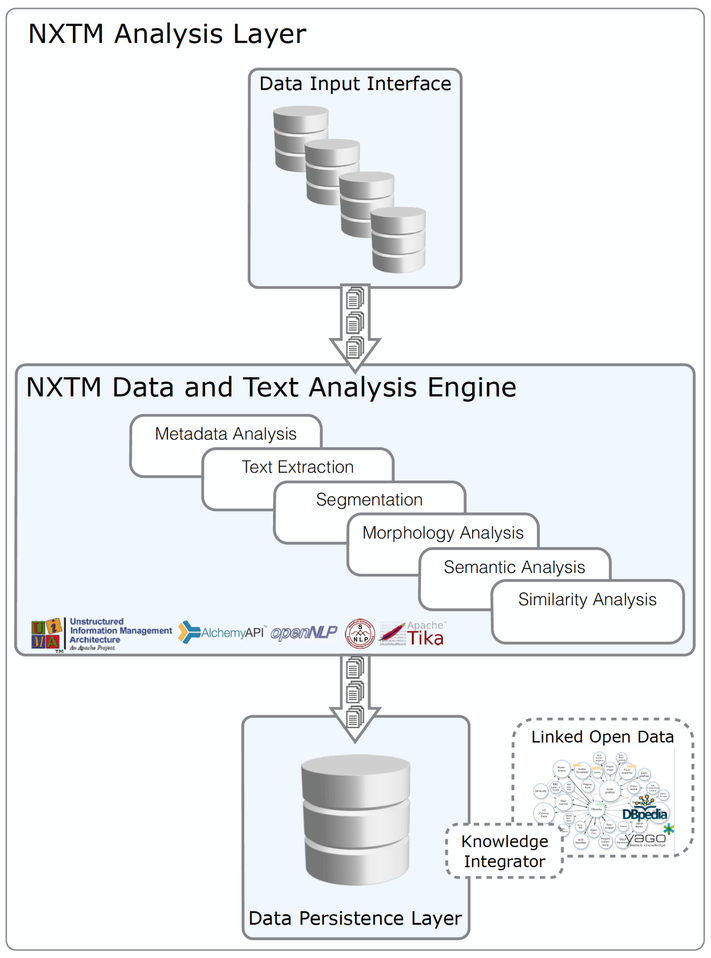
Implementation

Data input interface

- pipeline of documents to be analysed, updated or removed from the system
- providing data from many different sources

NLP analysis

1. import of documents to be analysed from the input pipeline
2. language identification, MIME-Type and metadata analysis
3. natural language processing in chained analysis engines and annotating semantic information
4. similarity calculation and document clustring
5. storing the documents and extracted data in a database, updating the search index
6. mapping annotated entities and their attributes with LOD knowledge databases



Clustering

- during the data analysis the documents are additionally clustered according to their simliarity
- in a hierarchical cluster it is possible to quickly find the most simliar document to the searched phrase
- similiarity measure is used to calulate distances between nodes in the result graph, reflecting the relevance between documents

LOD Integration

- extracted semantic information (i.e. entites, their attributes and relationships) may build a local knowledge graph
- by using web interfaces such a graph can be enhanced through integration with existing online knowledge databases and ontologies (Linked Open Data)

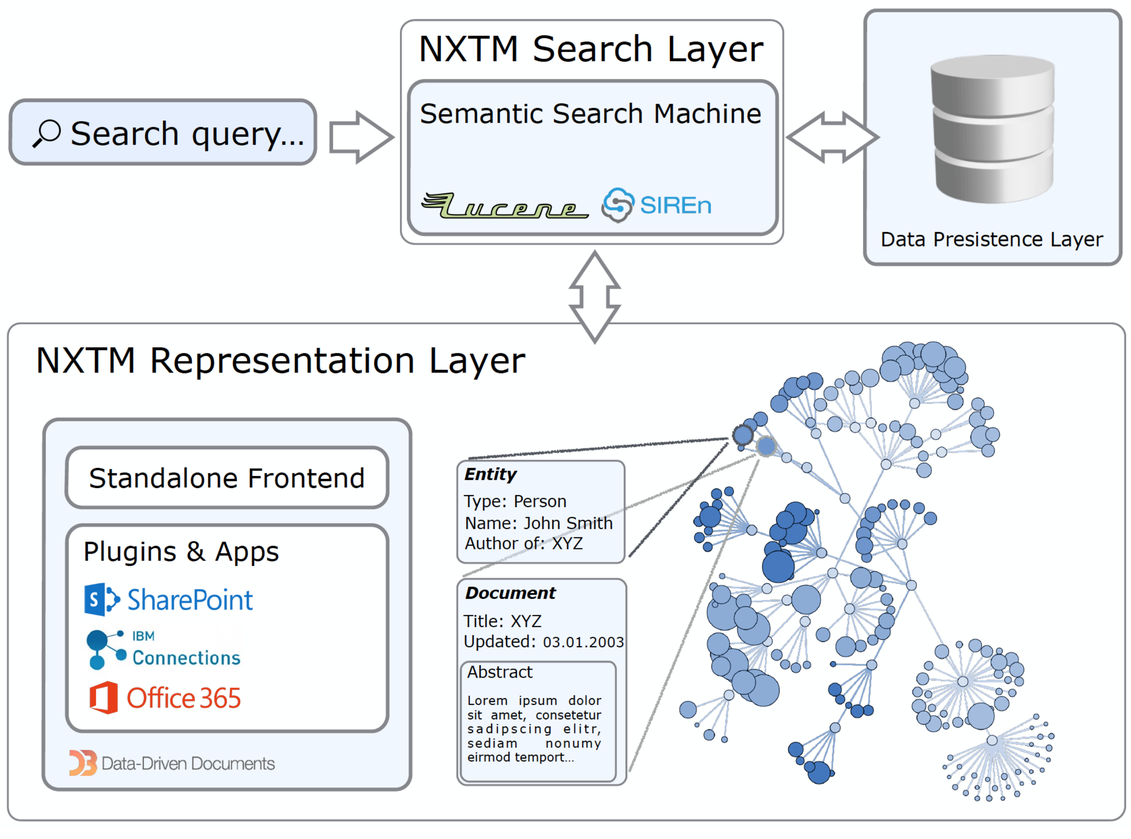
GUI Representation

Representation Layer

- search results are represented as an interactive graph with nodes and edges - a network of related documents, entities and metadata
- real time browsing of the graph enables the user to discover other relevant sources of information and their dependencies
- available as a standalone frontend as well as a plugin for other information management platforms

Semantic Search

- direct queries to a DB for retrieving the analysed data is an inefficient way of searching information:
  - user is unaware of the ontology of the persisted data
  - requires knowledge of a query language (SQL, SPARQL)
- a semantic search machine can effectively search for linked and hierarchical data, regardless to the data schema used



Partners/Cooperations

Hochschule Zittau/Görlitz
UNIVERSITY OF APPLIED SCIENCES

GABO®
Potenziale entfalten.

ZIM
Zentrales Innovationsprogramm Mittelstand

Bundesministerium für Wirtschaft und Energie

STUDYING_WITHOUT_BORDERS