Meta-X: Metadata Knowledge Discovery for Context Aware Business Intelligence

Hiba Khalid
Supervisor: Esteban Zimanyi
Co-Supervisor: Robert Wrembel
CPC Chair: Oscar Romero
Presentation Outline

• Introduction
• Literature Review (State of the Art)
• Problem Statement
• Methodology
• Progress Review
  • Project Planning
  • Research Dissemination Planning
• Conclusion
Introduction

Dataset Collection Explosion
Many organizations face the problem of integrating **multiple data sources** to attain business intelligence (BI)

Lack of Integration Principle
The BI teams do not have a generalized framework for dataset integration or processing

Data Lakes & Freedom
The problem is typically addressed using the **no uniformity** and freedom provided by **data lakes**. Data lakes have no uniformity and that’s a problem.

Data Fishing From Lakes
The next challenge is to fish or **retrieve the right dataset** for the corresponding problem, process or operation.
Problem Statement

The integration of independent data sources using metadata as knowledge base of meta-learning in order to obtain enhanced business intelligence.

Defined Goals

G1: Integrate independent data sources
G2: Construct knowledge base using metadata
G3: Minimize number of fetch requests in data lakes
G4: Increase accuracy of business analytics using deep learned metadata
Metadata Redefined-I

Metadata Classes/Groups

The traditional metadata does not provide conformity to increasing data integration and analytic processes. In order to attain power over data itself scientists have concluded a redefined class of metadata based on context and domain of application operation.

<table>
<thead>
<tr>
<th>Metadata Groups</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Size, formats, aliases, last modified time, access, control lists</td>
</tr>
<tr>
<td>Content Based</td>
<td>Schema, number of records, data fingerprints, key fields, frequent data tokens, similar datasets</td>
</tr>
<tr>
<td>Provenance</td>
<td>Reading jobs, writing jobs, downstream datasets, upstream datasets</td>
</tr>
<tr>
<td>User Supplied</td>
<td>Descriptions, annotations</td>
</tr>
<tr>
<td>Team and Project</td>
<td>Project description, owner, team name</td>
</tr>
<tr>
<td>Temporal</td>
<td>Change history</td>
</tr>
</tbody>
</table>

# Metadata Redefined-II

## Metadata at Use

### Descriptive Metadata
- **Created By:** Lee
- **Date of Creation:** 13 JUNE 2012
- **Version:** 2.0
- **Subject:** Member listing

- For discovery of data
- For displaying data such as transactional data (OLTP)
- For interoperability

### Structural Metadata
- **Table Index:** 102
- **P-key:** Yes
- **F-key:** Yes
- **Interdependencies:** 104, 106, 109

- Navigation and presentation
- Internal structure + relationship description
- Foreign Key CR-ID was included in table 102

### Administrative Metadata
- **Data Type:** Integer, real, Text
- **Access Rights:** Admin Only
- **Data Migration:** Yes
- **Last Migration:** 10 January 2012

- Short-term and long-term management processing
- Technical data: creation, quality control, rights management, access control

---

The research is based on aggregated literature in field of heterogeneous data sources, data lakes, integration of data sources, metadata extraction, enrichment, profiling.

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>Details</th>
<th>Drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOODS: Organizing Google’s Datasets</td>
<td>Metadata classification, enrichment, logging, cataloging, Provenance enrichment, enhancement of metadata</td>
<td>Not designed for common business analytics in companies. Costly Requires domain altering Not generic</td>
</tr>
<tr>
<td>Evaluation of Metadata Representations in RDF stores</td>
<td>Model for storing metadata along RDF data for big data at implementation and conceptual level.</td>
<td>Does not handle the cost of querying disintegrated data.</td>
</tr>
<tr>
<td>Web Tables: Exploring the Power of Tables on the Web</td>
<td>The identification and extraction of labelled schema as structured datasets. Analyzes and answers the query of traversing structured web tables in search engines. Schema auto complete. Attribute co-relations etc.</td>
<td>Only works for web tables and structured data. Auto-Complete schemas but no correlation between disintegrated schema representations.</td>
</tr>
</tbody>
</table>
Research Scenario: Credit Risk Modelling

**Information**
- Stock purchase: 2013; Name: ‘John Smith’
- Tennis Match Ticket: 2014; Purchase Type: Online; Payment Status: Paid
- Online Shopping: 2015; Purchase Mode: Credit: Online; Payment Status: Unpaid; Name: John

**Possible Queries**
Q1: Will John smith pay her credit card bill by the end of 2015?
Q2: Is John smith and John the same entities?
Research Explication

1. Metadata Collection
   To gather metadata of existing data, extract it and clean it for further processing

2. Deep Metadata Layer
   To run the metadata through deep learning network and create a ‘Deep self actualized metadata layer’

3. Process Answering
   To use this MD library for answering business and customer queries

4. Reinforcement
   To Reinforce the Membrane using Feedback Loop in the form of rewards and failures
Research Methodology

Process Details: Types of metadata formats for collection

Example Metadata to be extracted includes: Title, author, date, creator, about, subject etc.

RDF Schema Specifications

1: The metadata can be collected in different forms such as xml, RDF representation, even log files, schema etc.
Research Methodology

Process Details: Metadata Collection & Enhancement

2: The collection leads into metadata classes and categories for assignment

3: The third step focuses on enhancement of collected and categorized metadata
# Metadata Visualization Process

## Metadata

<table>
<thead>
<tr>
<th>Size</th>
<th>Provenance</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Written by: Task A</td>
<td>Credit.nlu.BNP</td>
</tr>
<tr>
<td>400</td>
<td>Read by: Reader A</td>
<td>Behaviour.anl.Schema</td>
</tr>
<tr>
<td>800</td>
<td>Corrected by: Task A</td>
<td>.....</td>
</tr>
<tr>
<td>600</td>
<td>Updated by: Task B</td>
<td>.....</td>
</tr>
<tr>
<td>300</td>
<td>Trained: Task D</td>
<td>.....</td>
</tr>
</tbody>
</table>
Deep Learning

• ANN’s but better!
• Multiple Hidden Layers
• Supervised, Unsupervised or Semi-supervised
• Learning Data Representations

• Automatic feature detection
• Hierarchical feature learning
• Multiple level representation
Learning Visualization: Learning to Predict Customer Behavior

Customer Data/History
Credit Relevant, personal etc.
DATA(Credit-Loan, Question)
Question: Is John capable of paying back full loan in next 3 years?

Deep Network

Input Layer

Output

HL 1

HL 2

HL 3

HL 4

John Smith Cannot pay loan by 2021

HL 1: John Smith

HL 2: Cannot pay

HL 4: By 2021

HL 3: Pay debt
Learning Visualization: Feedback & Reinforcement

RL offers the algorithms to learn from its own actions of classifications, predictions and decisions.
Research Inputs to Outputs Mapping

**Inputs:**
- Metadata
- Search Queries
- Ontologies
- Feedbacks

**Outputs:**
- Enriched Metadata knowledge base
- Predictive Analytics
- Entity Relations
- Agent Awareness
Research Limitations & Constraints

**Restrictions & Constraints**
- Metadata Enrichment
- Profiling
- Limited independent data sources (4)

**Limitations of Project**
- Does not handle ontology alignment
- Does not handle ontology evolution
- Does not contain aggregated datasets
## Summary & Conclusion

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Description</th>
</tr>
</thead>
</table>
| Milestones Achieved           | Literature Review  
Client Side Query Analysis (Text Based)  
Research Article-I based on scientific study variables and associated dependencies.  
Research Article-II is based on the construction of data composites for metadata discovery |
| Milestones Planned            | Metadata Extraction  
Metadata Enrichment  
Metadata Classification & Learning |
## ECTS Planned

<table>
<thead>
<tr>
<th>Activity</th>
<th>Place</th>
<th>ECTS</th>
<th>G/I/PC</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>January 2017 to December 2017</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer School</td>
<td>Brussels</td>
<td>2</td>
<td>PC</td>
<td>In Progress</td>
</tr>
<tr>
<td>French Language Course</td>
<td>Brussels</td>
<td>1.5</td>
<td>G</td>
<td>Planned</td>
</tr>
<tr>
<td>Research Publishing</td>
<td>Brussels</td>
<td>2</td>
<td>I</td>
<td>Planned</td>
</tr>
<tr>
<td>OPEN HPI Semantic Web Course</td>
<td>Online</td>
<td>1</td>
<td>PC</td>
<td>In Progress</td>
</tr>
<tr>
<td><strong>January to December 2018</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Development Studio</td>
<td>Poznan</td>
<td>3</td>
<td>PC</td>
<td>Planned</td>
</tr>
<tr>
<td>Technical &amp; Scientific Writing</td>
<td>Poznan</td>
<td>3</td>
<td>G</td>
<td>Planned</td>
</tr>
<tr>
<td>Data Mining and Analysis</td>
<td>Poznan</td>
<td>5</td>
<td>PC</td>
<td>Planned</td>
</tr>
<tr>
<td>Polish Language Course</td>
<td>Poznan</td>
<td>1.5</td>
<td>G</td>
<td>Planned</td>
</tr>
<tr>
<td><strong>January 2019 to January 2020</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Seminar</td>
<td>Brussels</td>
<td>1</td>
<td>PC</td>
<td>Planned</td>
</tr>
<tr>
<td>Internship</td>
<td>Brussels</td>
<td>2</td>
<td>PC</td>
<td>Planned</td>
</tr>
</tbody>
</table>
# PHD Timeline

<table>
<thead>
<tr>
<th>Sessions</th>
<th>Details</th>
<th>Status</th>
</tr>
</thead>
</table>
| Spring 2017 [Jan-July] | Research methodology Construction (Done)  
Literature Review (Done)  
Submission of DPP (Done)  
**Planned Publications:**  
(Completed, Reviewed, ready for Submission)  
b. Deep metadata Knowledge graphs for Semantic Web & Business Intelligence. | Completed   |
| Fall 2017-2018[Sept-Feb] | Moving to host university (PUT)  
Submission of Conference Paper-II (In Progress)  
**Planned Publications:**  
a. Understanding the metadata modelling in semantic business;  
b. Conceptual meta-data to automated metadata. | Planned     |
| Spring 2018 [Mar-July] | Submission of TPR  
Submission of Journal-I: SEMX: A learning model of metadata for high speed semantic knowledge | Planned     |
| Fall 2018-2019[Jan-July] | Proof of concept-I  
Submission of Journal II: Learning as a reinforced activity: Meta Composites in Business Intelligence | Planned     |
| Fall 2019-2020 [Jan-July] | Thesis write up and Defense | Planned }
References

- Vanden Berghen Frank, Q-Learning, IRIDIA, Universit Libre de Bruxelles.