Implementation of an exploratory workbench for identifying similar design decisions

Prateek Bagrecha, Garching, 12.02.2018, Advisor: Manoj Mahabaleshwar

Software Engineering betrieblicher Informationssysteme (sebis)
Fakultät für Informatik
Technische Universität München

wwwmatthes.in.tum.de

Agenda

❑ Introduction
❑ Research Questions
❑ Requirements
❑ System Design
❑ Process Overview
❑ Implementation Overview
❑ Evaluation
❑ Lessons Learned

In software architecture, *Design Decisions* are **decisions that *address architecturally significant requirements*.** They are

❑ Hard to make

❑ Costly to Change

❑ ***Often influence similar concerns ➔ Reuse ?***

Could knowledge about past decisions be used to make new informed decisions ?

Example: Comparing Two Decisions

| Issues	SPARK-8321	SPARK-19625
Description	Authorization Support(on all operations not only DDL) in Spark Sql	Authorization Support(on all operations not only DDL) in Spark Sql version 2.1.0
Concepts	Apache, SQL, authentication	Apache, SQL, authentication
Keywords	Spark, operations, Support, Authorization	Spark, operations, Support, Authorization
Components	SQL	Spark Core, SQL
Issue Type	Improvement	Improvement
Created	12/Jun/15 03:34	16/Feb/17 09:36
Resolved	16/Jun/16 08:22	24/Mar/17 01:21

Research Questions

❑ What are the functional and non-functional requirements of a workbench that supports identifying similar design decisions?

❑ How to identify similar design decisions using context-aware similarity measures and clustering analysis?

❑ How can a workbench support end-users in identifying the contextual parameters that are necessary for identifying similar design decisions ?

Requirements

❑ Workbench to provide both UI & Restful APIs

    ❑ for creating & configuring experiments for clustering design decisions

    ❑ to input new design concern and predict similar past design decisions

❑ Workbench shall abstract all operations related to identifying similar design decisions

❑ Automated import of data from SocioCortex & Amelie knowledge base

❑ Import different data formats

❑ Extension points for using multiple machine learning libraries

❑ Workbench is extensible without significant impact to system design
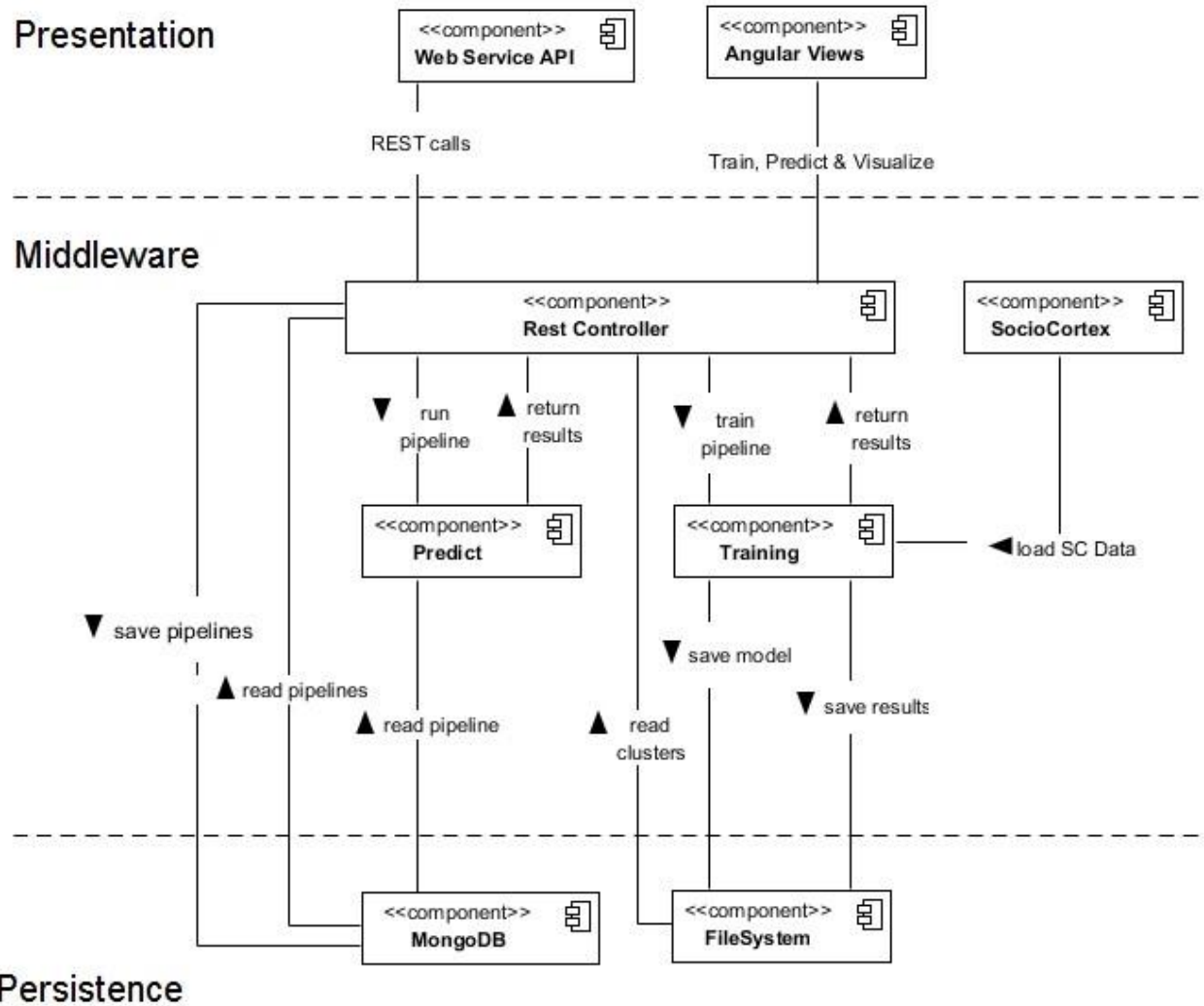
Powered by yFiles
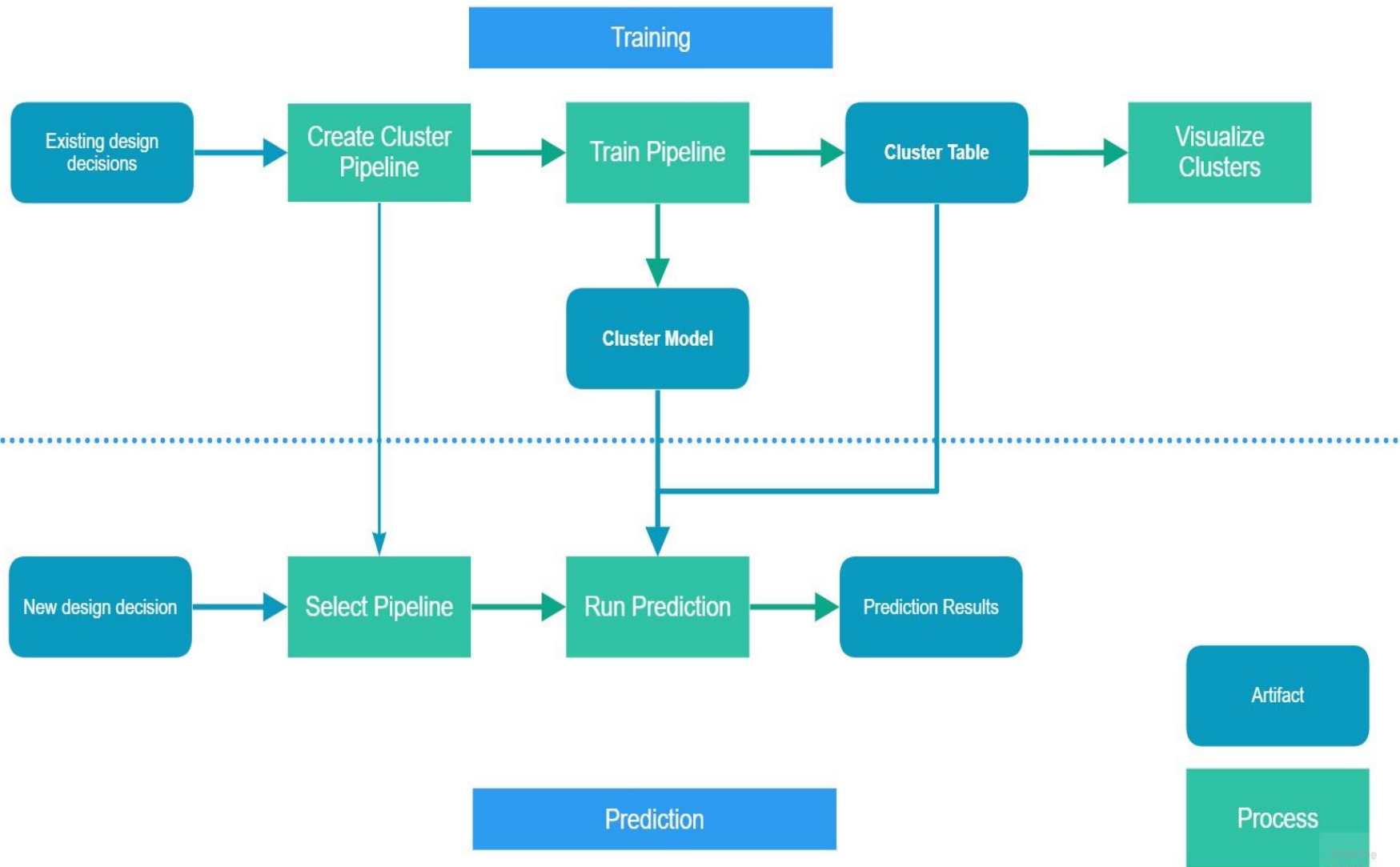
DEMO

Evaluation Strategy

Evaluation Steps

❑ For each design decision from the training dataset

    ❑ Mark related design decisions (related to, parent tasks, duplicates etc.)

    ❑ Run Predict Pipeline

    ❑ Check if the returned results contains related design decision

Datasets

❑ 2 Open Source Projects : Apache Solr & Hadoop Common

❑ 1 Component Based Cross Project Decisions

Evaluation Results

Project	Doc Key	Sample Results	Cosine Similarity	Jaccard Similarity	Duplicate	Related (Related To/Part of/Depends on)
Solr	SOLR-236	SOLR-1311	99.83	10.18	No	Yes
		SOLR-237	99.42	26.51	No	Yes
Solr	SOLR-373	SOLR-7986	99.41	60.00	Yes	No
SQL Component	CARBONDATA-440	CARBONDATA-504	93.20	28.57	No	Yes
		CARBONDATA-503	93.20	28.57	No	Yes

❑ Duplicates with relatively high cosine similarity and Jaccard similarity
❑ Related issues (related to, sub-tasks, duplicated by, parent etc.)

Industrial Impacts

❑ Connected Mobility Lab, Siemens
   ❑ Do not maintain related issues
❑ Digital Factory- Motion Control, Siemens
   ❑ Expert Recommendation

Lessons Learnt

❑ No two machine learning libraries are the same

   ❑ Different representation of ml models

   ❑ Different representation of results

❑ Occurrence of distinct decisions in the same cluster ➜ Model Tuning & Retraining

❑ Low number of related design decisions across projects

❑ Inability to recognize some related words for example: upsert is related to update.

Thank you

Future Works

❑ Implement cross functional pipelines ➜ working with different libraries within a single pipeline

❑ Custom implementation of clustering algorithm that supports cosine similarity as distance measure

❑ Support soft clustering mechanism

❑ Pipeline retraining & model tuning

❑ Extended visualization of results corresponding to clustering algorithms

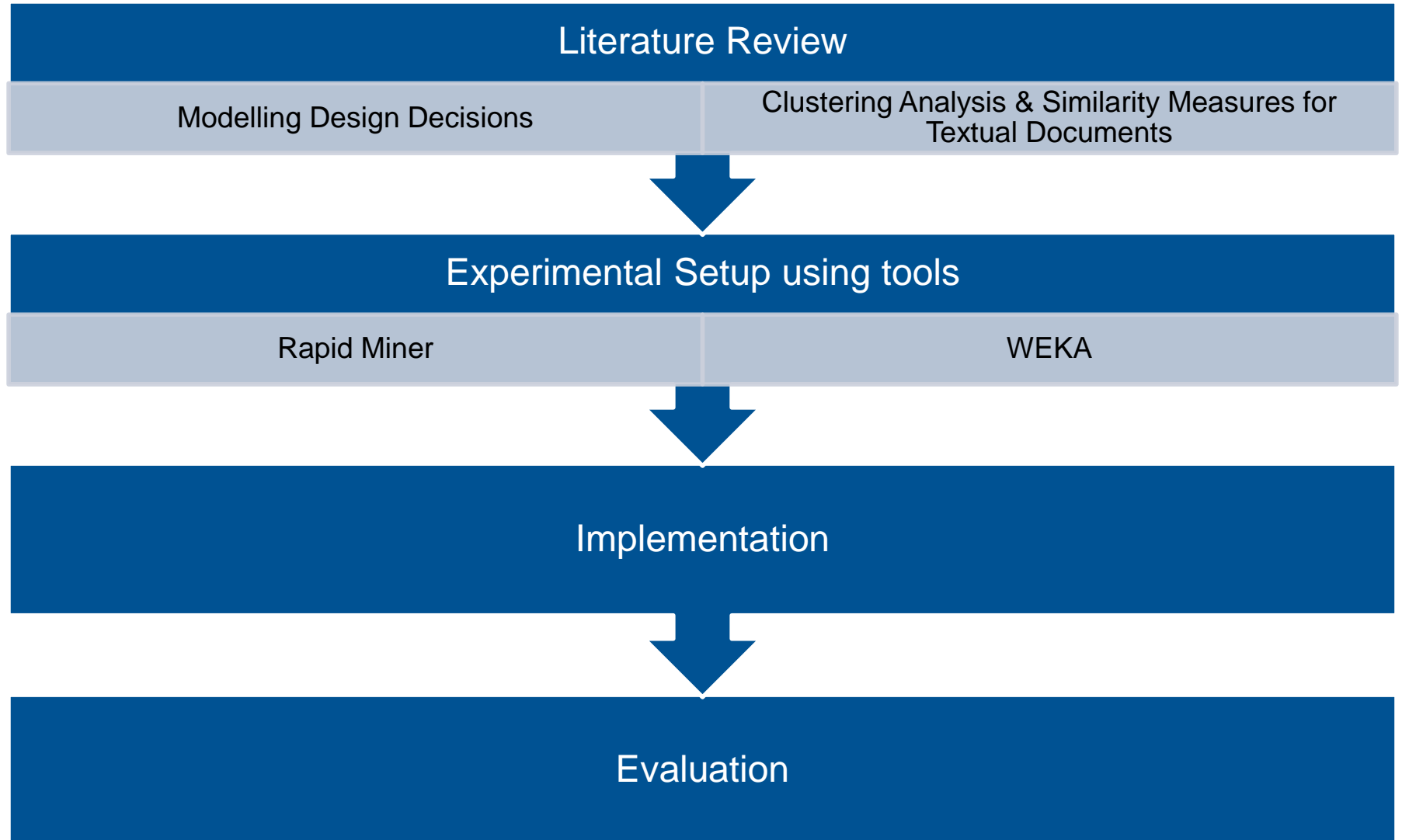❑ Further evaluation of the workbench

Evaluation Results
Performance

Project	Document Size (KB)	Members	Training Time	# of Clusters	Average cluster size
SocioCortex	603	726	18.93s	20	37
Apache Solr	6411	6175	1.2mins	30	206
Hadoop Commons	4024	6262	46.55s	20	313
SQL Component	14107	10069	1.8	30	334

Identifying similarities in design decisions (for an organization)

Motivation

By reusing knowledge from past decisions

❑ Documentation - specifying constraints on similar design decisions

❑ Communication - visual representation of related design decisions

❑ Complexity - Inferring the complexity for addressing similar design decisions

Literature Review

| Modelling Design Decisions | Clustering Analysis & Similarity Measures for Textual Documents |

Experimental Setup using tools

| Rapid Miner | WEKA |

Implementation

Evaluation

Helpful if second reporter could have been informed about the similar design decision made in past

❑ Reduced time to analyse
❑ Reduced time to resolution
❑ Reduced turn-around time for expert feedback

Given an open design decision, search the knowledge base for similar earlier made design decisions.