Julius-Maximilians-
UNIVERSITÄT WÜRZBURG

Universitätsklinikum Würzburg

Deutsches Zentrum für Herzinsuffizienz Würzburg

Leon Liman, Georg Fette, Georg Dietrich, Frank Puppe

Improving execution speed of i2b2 queries by using a Solr index based query engine

Introduction

i2b2 [1] is a popular data warehouse framework for the support of clinical research, e.g. by facilitating the estimation of cohort sizes for planned studies. Unfortunately the execution speed of the standard i2b2 query engine is slow depending on the amount and types of the selected query attributes as well as their combination in the query.

Solr [2]

- Based on Apache Lucene [3] and therefore with the same scalable, high-performance and near real-time indexing
- Optimized for high volume traffic
- Standards based open interfaces (e.g. JSON)
- Comprehensive administration interfaces built-in
- Advanced search capabilities

Solr index based query engine (SIBQE)

- Created as an alternative to the standard i2b2 query engine
- The Solr index is loaded with data from the Clinical Research Chart (CRC) and Ontology Management (ONT) databases of an existing i2b2 installation.
- SIBQE provides REST-interfaces matching the ones provided by the standard i2b2 CRC- and ONT-cells so it can be transparently used by i2b2 Web Clients, Workbenches and any other clients relying on these interfaces.

Solr index creation

CRC data

i2b2 OBSERVATION FACT	
PATIENT_NUM	INT
CONCEPT_CD	VARCHAR
TVAL_CHAR	VARCHAR
NVAL_NUM	DECIMAL
OBSERVATION_BLOB	VARCHAR
ENCOUNTER_NUM	INT
START_DATE	DATETIME

i2b2 OBSERVATION FACT

i2b2 OBSERVATION FACT

...

Patient Solr/Lucene document	
PatientID	LONG
ValuesForConcept1	TEXT [ ]
EncounterIDsForConcept1	LONG [ ]
StartDatesForConcept1	DATE [ ]
ValuesForConcept2	FLOAT [ ]
EncounterIDsForConcept2	LONG [ ]
StartDatesForConcept2	DATE [ ]
...	

- All observation facts for one patient are merged into a single document.
- For every single concept of a patient all the concept's values, encounter IDs and start dates are saved in arrays.
- Similar documents are created for every encounter. The only difference of these documents is, that all the EncounterIDsForConceptX fields are replaced by a single EncounterID field. This is necessary to support "Occurs in Same Encounter"-constraints of i2b2's query language.

ONT data

i2b2 ONTOLOGY CONCEPT	
C_NAME	VARCHAR
C_FULLNAME	VARCHAR
C_BASECODE	VARCHAR
C_TOTALNUM	INT
C_METADATAXML	VARCHAR

Concept Solr/Lucene document	
Name	TEXT
ID	TEXT
ParentID	INT
ExternalID	TEXT
Project	TEXT
PatientCount	INT
DataType	TEXT
LowBound	FLOAT
HighBound	FLOAT

- For each i2b2 Ontology concept a matching Solr/Lucene document is created.
- The Information contained in i2b2's C_FULLNAME is saved as a single numeric ID (with a fixed prefix string) for each concept together with a numeric ParentID. The ParentID matches the numeric ID of the concept being the current concept's direct predecessor in the concept hierarchy.
- The C_BASECODE is splitted into the code system (saved as Project) and the actual code (saved as ExternalID).
- If present the data type and Low-/HighOfToxicValue are extracted from the C_METADATAXML field.

Docker-Image

- A Docker-Image is provided [4] containing all the necessary components (configured Solr-Server, Tomcat-Server, Java programs controlling the data flow, Apache Webserver hosting an preconfigured i2b2 Web Client).
- Only the credentials to the CRC- and ONT- i2b2 databases have to be provided as Docker environment variables.
- After data transfer the i2b2 PM cell has to be updated to point to the alternative CRC- and ONT-REST-Interfaces and the SIBQE is ready to be used by i2b2 clients for patient count requests.

Current Limitations

- Only supporting i2b2 databases stored on a Microsoft SQL Server [5]
- Only supporting patient count requests
- The SIBQUE-REST-Interfaces have limited capabilities compared to the i2b2 ones.

Results

The performance of the SIBQUE compared to the i2b2 standard query engine had been evaluated on different datasets with multiple queries of different complexity.

Dataset	i2b2 database size in GB	Solr index size in GB	Speed improvement
10.000 artificially generated patients	1,7	1	on average by a factor of 10 to 30
100.000 artificially generated patients	15	8	
1.000.000 artificially generated patients	166	97	
Real clinical data of about 1.000.000 patients	640	275	reaches factor 80 with certain query types

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

[1] Partners Healthcare. i2b2 software and documentation. 2017. https://www.i2b2.org/software. Accessed 21. September 2017
[2] Apache Software Foundation. Apache Solr. 2017. http://lucene.apache.org/solr. Accessed 21. September 2017
[3] Apache Software Foundation. Apache Lucene. 2017. http://lucene.apache.org. Accessed 21. September 2017
[4] Leon Liman (University of Wuerzburg). i2b2-Solr-Backend. 2017. https://hub.docker.com/r/uniwue/i2b2-solr. Accessed 21. September 2017
[5] Microsoft Corporation. Microsoft SQL Server. 2017. https://www.microsoft.com/sql-server. Accessed 21. September 2017