Efficient Keyword Search For The QLever SPARQL Engine

...or how we improved the QLever text search.

Nick Göckel
What is SPARQL?

- query language
- used for querying RDF data
  - common format to store data
  - uses only triples
Example query:

```
SELECT ?scientist WHERE {
  ?scientist <Award_Won> <Nobel_Prize_in_Physics> .
}
```

Run on database:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Albert_Einstein&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Physics&gt;</td>
</tr>
<tr>
<td>&lt;Carl_Bosch&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Chemistry&gt;</td>
</tr>
<tr>
<td>&lt;Charles_Darwin&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Royal_Medal&gt;</td>
</tr>
<tr>
<td>&lt;Marie_Curie&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Chemistry&gt;</td>
</tr>
<tr>
<td>&lt;Marie_Curie&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Physics&gt;</td>
</tr>
</tbody>
</table>
Example query:

```
SELECT ?scientist WHERE {
  ?scientist <Award_Won> <Nobel_Prize_in_Physics> .
}
```

Run on database:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Albert_Einstein&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Physics&gt;</td>
</tr>
<tr>
<td>&lt;Carl_Bosch&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Chemistry&gt;</td>
</tr>
<tr>
<td>&lt;Charles_Darwin&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Royal_Medal&gt;</td>
</tr>
<tr>
<td>&lt;Marie_Curie&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Chemistry&gt;</td>
</tr>
<tr>
<td>&lt;Marie_Curie&gt;</td>
<td>&lt;Award_Won&gt;</td>
<td>&lt;Nobel_Prize_in_Physics&gt;</td>
</tr>
</tbody>
</table>

Has result:

<table>
<thead>
<tr>
<th>?scientist</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Albert_Einstein&gt;</td>
</tr>
<tr>
<td>&lt;Marie_Curie&gt;</td>
</tr>
</tbody>
</table>
What is QLever?

- **SPARQL query engine**
  - runs SPARQL queries on RDF data bases
- **allows for combined search**
  - search on structural data from an RDF knowledge graph
  - but also on textual information from a collection of texts
# PREFIX lines can be ignored.

PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ql: <http://qlever.cs.uni-freiburg.de/builtin-functions/>

SELECT ?name ?text WHERE {
  ?scientist wdt:P166 wd:Q38104 . # wdt:P166 is "award received" and wd:Q38104 is "Nobel Prize in Physics".
  ?scientist rdfs:label ?name . # So we get "Albert Einstein" instead of Q937 as a result.
  ?text ql:contains-word "astrophysics" .
  FILTER (LANG(?name) = "en") # Filter out non-English results.
}

TEXTLIMIT 2

https://qlever.cs.uni-freiburg.de/wikidata/v42wB4
How did we improve the QLever text search?

1. Implemented new feature, that extends the text search
2. Improved the code structure of the text search
<table>
<thead>
<tr>
<th>Name</th>
<th>Text</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>John C. Mather</td>
<td>John Cromwell Mather (born August 7, 1946, Roanoke, Virginia) is an American astrophysicist, cosmologist and Nobel Prize in Physics laureate for his work on the Cosmic Background Explorer Satellite (COBE) with George Smoot.</td>
<td>astrophysicist</td>
</tr>
<tr>
<td>John C. Mather</td>
<td>Mather is a senior astrophysicist at the NASA Goddard Space Flight Center (GSFC) in Maryland and adjunct professor of physics at the University of Maryland College of Computer, Mathematical, and Natural Sciences.</td>
<td>astrophysicist</td>
</tr>
<tr>
<td>Vitaly Ginzburg</td>
<td>He also headed the Academic Department of Physics and Astrophysics Problems, which Ginzburg founded at the Moscow Institute of Physics and Technology in 1968.</td>
<td>astrophysics</td>
</tr>
<tr>
<td>Vitaly Ginzburg</td>
<td>Soviet astrophysicist Vitaly Ginzburg said that ideologically the &quot;Bolshevik communists were not merely atheists, but, according to Lenin's terminology, militant atheists&quot; in excluding religion from the social mainstream, from education and from government.</td>
<td>astrophysicist</td>
</tr>
<tr>
<td>Adam Riess</td>
<td>Adam Guy Riess (born December 16, 1969) is an American astrophysicist and Bloomberg Distinguished Professor at Johns Hopkins University and the Space Telescope Science Institute.</td>
<td>astrophysicist</td>
</tr>
<tr>
<td>Adam Riess</td>
<td>In astrophysics, Press is best known for his discovery, with Paul Schechter, of the Press–Schechter formalism, which predicts the distribution of masses of galaxies in the Universe; and for his work with Adam Riess and Robert Kirshner on the calibration of distant supernovas as &quot;standard candles&quot;.</td>
<td>astrophysics</td>
</tr>
<tr>
<td>Saul Perlmutter</td>
<td>The newest supercomputer Perlmutter, is named after Saul Perlmutter, an astrophysicist at Berkeley Lab who shared the 2011 Nobel Prize in Physics for his contributions to research showing that the expansion of</td>
<td>astrophysicist</td>
</tr>
</tbody>
</table>
Why do we want this feature?

- we can use information like normal variable
  - group by
  - join
  - filter
  - etc.
- allows user to run new set of queries
  - for example: https://qlever.cs.uni-freiburg.de/wikidata/onBH1f
Corona-Satzung Uni Freiburg
staatenloser Studienbe
Abweichend von § 9 Absatz 1 der Universität Freiburg (ZImmO) der Universität Freiburg ausländischen oder staatenlosen Personen Erscheinen und auf die Vorzüge bleibt nicht zugeschrieben, wenn er/sie bis zum 31. Dezember 2021 beziehungsweise in einem Bundesgebiet einreist. Nach seiner/ihrer Einreise in das Bundesgebiet bestehende Krankenkasse bestehende oder über eine Aufenthaltserlaubnis unverzüglich, ist der/die St

MSc Rahmenprüfungs-Studienleistungen und (1) Studienzeiten Studien-
Query:

```sql
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ql: <http://qlever.cs.uni-freiburg.de/builtin-functions/>

SELECT * WHERE {
  ?scientist wdt:P166 wdt:Q38104 .
  ?text ql:contains-word "astrophysics" .
}

TEXTLIMIT 2
```
Why is this structure bad?

- hard to read
- hard to maintain
- code duplication (e.g. join operation)
- hard to implement new features
Query:

PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ql: <http://qlever.cs.uni-freiburg.de/builtin-functions/>

SELECT * WHERE {
  ?scientist wdt:P166 wdt:Q38104.
  ?text ql:contains-word "astrophysics".
}

TEXTLIMIT 2
Implementing feature

- only need to change code behind ql:contains-word
  - needs to read extra information from the text index
- other operations can stay the same

⇒ new text search structure made implementation way easier
Evaluation

Feature Implementation:
- implemented with full functionality as imagined
- high code quality

Restructuring:
- increased readability
- increased maintainability
- decreased duplication
- high code quality
Questions?