Problem 5: Query Publication Data from zbMATH

AI1SysProj 2021/2022

Topic:	Semantic Web
Due on:	March 15, 2022
Version from:	February 22, 2022

1 Task Summary

Translate a large dataset about mathematical publications into RDF triples and load them into a triplestore. Then use SPARQL queries to answer a set of questions.

2 The Dataset

zbMATH [ZBM] collects abstracts and reviews of papers in the area of mathematics and its aplications. Recently, a lot of their data was made publicy accessible via an API as well as a dataset [Pet+]. You can download the dataset from [ZBMd] (the relevant file is zbMathOpen_OAIPMH_int.xml.bz2).

The dataset consists of records. Each record corresponds to one publication. The following entries in metadata/oai_zb_preview:zbmath are of interest to us:

- zbmath:document_id is an identifier for the publication.
- zbmath:classifications/zbmath:classification lists the classifications of the publication using the Mathematics Subject Classification [MSC].
- zbmath:author_ids/zbmath:author_id lists identifiers for the authors.
- zbmath:keywords/zbmath:keyword lists keywords of the publication.

RDF requires that we use URIs to identify resources. These could be anything, but we will use the following URLs for this assignment:

Entry type	**Example value**	**Associated URL**
Document id	1448.68463	https://zbmath.org/?q=an%3A1448.68463
Classification	03B10	https://zbmath.org/classification/?q=cc%3A03B10
Author	kohlhase.michael	https://zbmath.org/authors/?q=ai%3Akohlhase.michael
Keyword	semantic web	https://zbmath.org/?q=ut%3Asemantic+web

3 Using Blazegraph

For this assignment, you should use Blazegraph [BG], an open source triplestore. That means that you will have to create RDF triples from the dataset described above and import them into Blazegraph. You should then solve the problems via SPARQL queries (without directly using the original dataset).

You can start Blazegraph with `java -jar blazegraph.jar`. It then runs on port 9999 (`http://localhost:9999/` gives you access to the blazegraph workbench). It also creates a journal file `blazegraph.jnl` to store the data.

You can update the data with a `POST` request to `http://localhost:9999/blazegraph/namespace/kb/sparql?update=XYZ` where `XYZ` is the (URL-encoded) update. For example, `DROP ALL` deletes all triples from the database and `LOAD <file:///path/to/triples.rdf>` loads data from a file.

Similarly, you can send a SPARQL query with a `GET` request to `http://localhost:9999/blazegraph/namespace/kb/sparql?query=XYZ` where `XYZ` is the (URL-encoded) SPARQL query.

4 Problems and Solution Format

Problems are provided in an XML file and the solutions should also be encoded as an XML file. You find the problems, along with example problems and solutions at [A5]. Listing 1 illustrates the structure of problems and solutions. The following subsections will describe the different problem types.

4.1 Problem type: coauthors

List all co-authors of someone. This should be fairly straight-forward once you have loaded the data into Blazegraph. Listing 2 shows an example.

4.2 Problem type: msc-intersection

List all publications that have a specific set of classifications. Listing 3 shows an example that looks for publications combining group theory (classification `20`) and knowledge representation in artificial intelligence (classification `68T30`). One of the publications combining those classifications is `5155089`, which has the classifications `68T30`, `20F10`, `68T37` and `68T05`.

```
<Problems>
  <Problem id="1" type="..." withquery="true">
    ...

  </Problem>
  <Problem id="2" type="..." withquery="false">
    ...
  </Problem>
  ...
</Problems>
```

```
<Solutions>
  <Solution id="1">
    ...
    <query>...</query>
  </Solution>
  <Solution id="2">
    ...
  </Solution>
  ...
</Solutions>
```

Listing 1: Example problems with their solutions. Every problem has an identifier that links it to the solution. The content of the problem and the solution depends on the problem type. For some problems, the SPARQL query that led to the solution should also be included.

```
  <Problem id="1" type="coauthors" withquery="true">
    <value type="of">https://zbmath.org/authors/?q=ai%3Aeinstein.albert</value>
  </Problem>
```

```
<Solution id="1">
  <value type="coauthor">https://zbmath.org/authors/?q=ai%3Aschrodinger.erwin</value>
  <value type="coauthor">https://zbmath.org/authors/?q=ai%3Aminkowski.hermann</value>
  ...
  <query>...</query>
</Solution>
```

Listing 2: Example problem and solution solution for problem type coauthors.

```
<Problem id="2" type="msc-intersection" withquery="true">
  <value type="msc">https://zbmath.org/classification/?q=cc%3A20</value>
  <value type="msc">https://zbmath.org/classification/?q=cc%3A68T30</value>
</Problem>

<Solution id="2">
  <value type="paper">https://zbmath.org?q=an%3A5155089</value>
  <value type="paper">https://zbmath.org?q=an%3A647709</value>
  ...
  <query>...</query>
</Solution>
```

Listing 3: Example problem solution solution for problem type msc-intersection.

Note that 20F10 is a subclassification of 20. It might make sense to add the subclassification relation to your dataset.

4.3 Problem type: top-3-keywords

List the three most common keywords of the publications of an author. There might be additional constraints on the publication year requiring that only consider publications before and/or after a particular year. Listing 4 shows an example.

4.4 Problem type: coauth-dist

Find a minimal connection between two authors X and Y. A connection between X and Y is a sequence of authors $a_0, \ldots, a_n$ where $a_0 = X$, $a_n = Y$ and for all $0 \leq i < n - 1$ there is a publication that has both a_i and a_{i+1} as authors.

The length n of such a connection is a generalization of the Erdős number [EN], which measures the co-author distance to mathematician Paul Erdős, who has collaborated with many people.

In addition to the authors, you should also list the publications connecting them. Listing 5 shows a minimal connection from Michael Kohlhase to Paul Erdős.

Solving this problem probably requires multiple SPARQL queries.

```
<Problem id="3" type="top-3-keywords" withquery="true">
  <value type="of">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
  <value type="after">1999</value>
  <value type="before">2012</value>
</Problem>

<Solution id="3">
  <value type="keyword">https://zbmath.org?q=ut%3Amathematical+knowledge+management</value>
  <value type="count">4</value>
  <value type="keyword">https://zbmath.org?q=ut%3Asemantics</value>
  <value type="count">2</value>
  <value type="keyword">https://zbmath.org?q=ut%3Acut+elimination</value>
  <value type="count">2</value>
  <query>...</query>
</Solution>
```

Listing 4: Example problem solution solution for problem type top-3-keywords.

```
  <Problem id="4" type="coauth-dist" withquery="false">
    <value type="from">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
    <value type="to">https://zbmath.org/authors/?q=ai%3Aerdos.pal</value>
  </Problem>

<Solution id="4">
  <value type="author">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
  <value type="paper">https://zbmath.org?q=an%3A5576982</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Adavenport.james-harold</value>
  <value type="paper">https://zbmath.org?q=an%3A3770941</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Aguy.michael-j-t</value>
  <value type="paper">https://zbmath.org?q=an%3A3643271</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Aerdos.pal</value>
</Solution>
```

Listing 5: Example problem solution solution for problem type coauth-dist.

5 Random Tips

- A key challenge is dealing with the large dataset file. One of the problems is that parsing the entire file into a DOM probably won't work for you. Instead, you could consider using a SAX parser. Another option might be to use the streaming capabilities of XSLT 3 (but I haven't tried that myself).
- Importing the RDF into Blazegraph takes time. It's probably a good idea to avoid doing that more often than necessary – e.g. by testing a small subset first.
- coauth-dist problems are a bit tricky because the queries might branch out too much. Consider controlling it a bit using FILTER.
- Have fun :-)

References

[A5] *Assignment 5: Semantic Web*. URL: https://gitlab.rrze.fau.de/wrv/AISysProj/ws2122/semweb/assignment.

[BG] *Welcome to Blazegraph*. URL: https://blazegraph.com/ (visited on 02/04/2022).

[EN] *Erdős number*. URL: https://en.wikipedia.org/wiki/Erd%C5%91s_number (visited on 02/07/2022).

[MSC] *Mathematics Subject Classification – MSC2020*. URL: https://zbmath.org/classification/ (visited on 02/04/2022).

[Pet+] Matteo Petrera et al. "zbMATH Open: API Solutions and Research Challenges". In: *Submitted to Joint Conference of Digital Libraries 2021 (JCDL '21), September 27–30, 2021, Online, Global*.

[ZBM] *zbMATH Open*. URL: https://zbmath.org (visited on 02/04/2022).

[ZBMd] *zbMATH Open Dataset*. URL: https://gl.kwarc.info/SIGMathLing/dataset-zbmath-open-2021 (visited on 02/04/2022).