Assignment 4: Query publication data from zbMATH

AI-1 Systems Project (Winter Semester 2023/2024)

Jan Frederik Schaefer

Friedrich-Alexander-Universität Erlangen-Nürnberg, Department Informatik

Topic:	Semantic web
Due on:	February 10, 2024
Version from:	December 20, 2023
Author:	Jan Frederik Schaefer

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. A link to the dataset is posted in the matrix room. This assignment **does not involve any data scraping**.

Didactic objectives

1. Get hands-on experience with RDF and SPARQL,
2. learn how to find information about standards like RDF and SPARQL,
3. get to know some of the challenges that come from working with large datasets.

Prerequisites and useful methods

1. The XML format and related technologies (in particular, SAX parsing might be useful for the large data set, and DOM parsing with XPaths for the problem files),
2. RDF, triple stores and SPARQL (the AI lecture introduces these concepts briefly, but there are also plenty of online resources),
3. the basics of HTTP requests (for communicating with the triple store – ask for help if you have problems with this),
4. URLs and percent-encoding (also known as URL encoding).

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+]. Furthermore, we have created a mini dataset, which contains only a small subset of the entries and might therefore be easier to work with. A download link to the datasets will be posted in the matrix room for this assignment. You do not have to do any data scraping.

The datasets consist 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.
- zbmath:publication_year is the year 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

You should translate the datasets into RDF triples and load them into a triple store so that you can run SPARQL queries (you have to solve the problems using SPARQL queries, not using the original dataset). There are many different triple stores and you are free to pick one that works for you. In my (limited) experience, the open source triplestore Blazegraph [BG] is relatively easy to set up and use.

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.

```
<Problems>
  <Problem id="0" type="...">
    ...

  </Problem>
  <Problem id="1" type="...">
    ...

  </Problem>
  ...
</Problems>
```

```
<Solutions>
  <Solution id="0">
    <Query>...</Query>
    ...
  </Solution>
  <Solution id="1">
    <Query>...</Query>
    ...
  </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. Every solution should also include the SPARQL query for solving it.

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. There are separate problem files for the mini dataset and the full dataset. You can find the problem files, along with example problems and solutions at [AR]. Listing 1 illustrates the structure of problems and solutions. Each problem should be answered with a single SPARQL query (and optionally some light post-processing). The following subsections will describe the different problem types.

```
<Problem id="0" type="coauthors">
  <Author>https://zbmath.org/authors/?q=ai%3Aeinstein.albert</Author>
</Problem>
```

```
<Solution id="0">
  <Query>...</Query>
  <Author>https://zbmath.org/authors/?q=ai%3Aschrodinger.erwin</Author>
  <Author>https://zbmath.org/authors/?q=ai%3Apauli.wolfgang</Author>
  ...
</Solution>
```

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

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. 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-authors

List the ten authors with the most publications with a particular keyword. Only papers published in a particular range of years should be considered (the limiting years should be excluded). Listing 4 shows an example.

5 Submission

At the deadline, we will download a snapshot of your repository. It should contain:

```
<Problem id="0" type="msc-intersection">
  <Classification>https://zbmath.org/classification/?q=cc%3A20</Classification>
  <Classification>https://zbmath.org/classification/?q=cc%3A68T30</Classification>
</Problem>

<Solution id="0">
  <Query>...</Query>
  <Paper>https://zbmath.org?q=an%3A5155089</Paper>
  <Paper>https://zbmath.org?q=an%3A647709</Paper>
  ...
</Solution>
```

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

```
<Problem id="0" type="top-authors">
  <Keyword>https://zbmath.org/?q=ut%3Aartificial+intelligence</Keyword>
  <AfterYear>1974</AfterYear>
  <BeforeYear>1981</BeforeYear>
</Problem>

<Solutions>
  <Solution id="0">
    <Query>...</Query>
    <Author count="6">https://zbmath.org/authors/?q=ai%3Alukasiewicz.thomas</Author>
    <Author count="3">https://zbmath.org/authors/?q=ai%3Ayager.ronald-r</Author>
    <Author count="3">https://zbmath.org/authors/?q=ai%3Astrauss.olivier</Author>
    ...
  </Solution>
</Solutions>
```

Listing 4: Example problem solution for problem type top-authors.

1. All your code.
2. Solution files (both for the mini dataset and the full dataset).
3. The RDF generated for the mini dataset (you can compress it to reduce the file size).
4. A README file explaining how to run your code.
5. A brief evaluation of your solution (what approach worked well, what did not?) either as a PDF file ($\approx$ 1 page) or as part of your README.md.

6 A Few Tips

- Make sure that you understand what RDF is.
- 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).
- Different formats for RDF. The most established is probably RDF-XML. You could also use the more human-readable Turtle syntax or N-Triples (a minimal subset of Turtle).
- Have fun :-)

7 Points

This is a relatively small assignment and only worth 80 points. You can get up to 20 points for the quality of the submission (README, evaluation, ...). Furthermore, you get 1 points for every correctly solved problem. There are 30 problems for the mini dataset and 30 problems for the entire dataset, which means that you can get a total of 60 points for the solutions. Alternatively, you can also get 2 points for every correctly solved problem for the entire dataset, and 0 points for the mini dataset so that you do not have to bother with the small dataset if you prefer. We will take the better of the two options (i.e., the one that gives you more points). Note that points will be subtracted if the solution files are not in the right format.

If the grading scheme doesn't seem to work well, we might adjust it later on (likely in your favor).

References

[AR] *Repository for Assignment 4: Query publication data from zbMATH*. URL: https://gitlab.rrze.fau.de/wrv/AISysProj/ws2324/a1.4-query-math-data/assignment.

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

[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).