ir_metadata: An Extensible Metadata Schema for IR Experiments

Timo Breuer, Jüri Keller, and Philipp Schaer

TH Köln - University of Applied Sciences, Germany, firstname.lastname@th-koeln.de

Motivations and contributions

- Shared task conferences (CLEF, NTCIR, TREC) archive **experimental artifacts**, i.e., run files
- Runs are a **valuable resource** for baselines and meta-evaluations but **the data does not provide context**
- Annotating run files with metadata information facilitates **better comparability, transparency, and reproducibility**

As a solution, we contribute:

- **Metadata schema** based on the **PRIMAD** taxonomy
- Software support by `repro_eval`
- **Open-access dataset**
- **Meta-evaluations** / reproducibility studies

Metadata schema and annotations

- The schema is based on **PRIMAD**, intended to be **extensible**, and we are open for **proposals of new metadata fields**.
- **PRIMAD** is a taxonomy that is based on the components that can affect the **reproducibility** of an IR experiment.
- We extend the taxonomy which allows a more detailed description of the experiments. The **project's website** provides **checklists** for each PRIMAD component.
- The annotations follow the **YAML syntax** and are added to the beginning of the TREC run files as a comment similar to a **file header** that tells us something about the rankings. `trec_eval` will support comments like these in future releases.

Example

```
# ir_metadata.start
# platform:
#   hardware: ...
#   operating system: ...
#   software: ...
# research goal:
#   venue: ...
#   publication: ...
#   evaluation: ...
# implementation:
#   executable: ...
#   source: ...
# method:
#   automatic: ...
#   indexing: ...
#   retrieval: ...
# actor:
#   name: ...
#   orcid: ...
#   team: ...
#   fields: ...
#   mail: ...
#   role: ...
#   degree: ...
# data:
#   test collection:
#     name: ...
#     source: ...
#     qrels: ...
#     topics: ...
#     ir_datasets: ...
# ir_metadata.end
307  Q0  497476  1  0.9931  bm25
307  Q0  469928  2  0.9674  bm25
307  Q0  125806  3  0.9623  bm25
307  Q0  504815  4  0.9453  bm25
307  Q0  392547  5  0.9223  bm25
...
```

Software support

Currently, the software supports ...

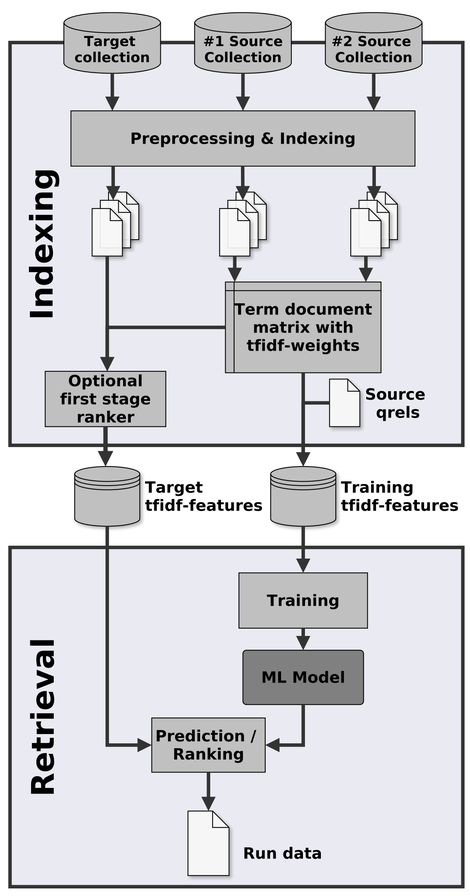
- ... the metadata **I/O handling** and automatic annotations of some components. The `MetadataHandler` can fetch information about the underlying platform and the implementation automatically.
- ... the **analysis of runs** with metadata annotations. The `MetadataAnalyzer` analyzes a directory that contains run files, and afterward, the `PrimadExperiment` evaluates the runs (examples are shown by the meta-evaluations).

All of the software features are added to `repro_eval`, which is a toolkit for reproducibility experiments. A **Google Colab notebook** exemplifies how the software and metadata can be used in your own implementations.

Dataset

We provide an annotated dataset that contains 463 run files. All of the runs are based on **cross-collection relevance feedback** as introduced by Grossman and Cormack as part of TREC Common Core in 2017 and 2018. Furthermore, we annotate **reimplementations** by Yu et al. (TREC, 2018; ECIR, 2019) and by us (SIGIR, 2020; CLEF, 2021). The dataset is hosted on **Zenodo** with DOI 10.5281/zenodo.5997491.

Cross-collection relevance feedback



1. **Derive tfidf representations** of documents for a given topic from one or two source collections
2. **Train a relevance classfier** with the tfidf representations and relevance labels
3. **Rank documents** of the target collection

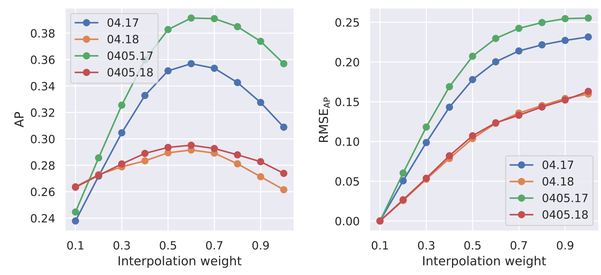
Dataset constellation

Actor	Method	Data	Runs
GC	GC'17	Core17	2
YXL		Robust04/05, Core17/18	327
BFFMSSS		Core17	100
GC	GC'18	Core18	2
BPS		Robust04/05, Core17/18	32

Meta-evaluations

We group the annotated runs into **three categories based on how the reimplementations relate to the original runs in terms of PRIMAD**. For instance, in the first experiment, we have just modified the method while keeping all of the other components fixed. In the second experiment, we varied all of the PRIMAD componets except for the data. And finally, in the third experiment, we varied all of the PRIMAD components.

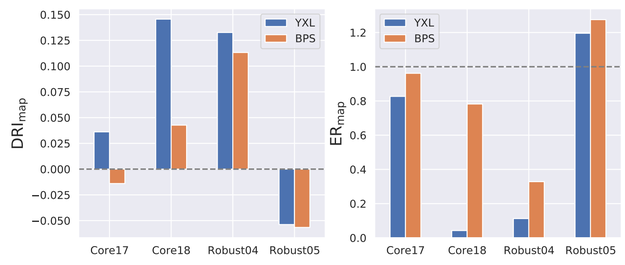
PRIM'AD: Parameter sweeps



P'R'I'M'A'D: Reproducibility analysis

Actor	GC	YXL	BFFMSSS
Baseline			
Average Precision	0.3711	0.4018	0.3612
Kendall's τ Union	1.0000	0.0086	0.0051
Rank-Biased Overlap	1.0000	0.1630	0.5747
Root Mean Square Error	0.0000	0.1911	0.1071
p-value	1.0000	0.1009	0.7885
Advanced			
Average Precision	0.4278	0.4487	0.4208
Kendall's τ Union	1.0000	0.0069	0.0111
Rank-Biased Overlap	1.0000	0.2231	0.6706
Root Mean Square Error	0.0000	0.2088	0.0712
p-value	1.0000	0.2785	0.8249
Overall effects			
Effect Ratio	1.0000	0.8267	1.0514
Δ Relative Improvement	0.0000	0.0362	-0.0123

P'R'I'M'A'D': Generalization



Resources

- **Website**
  `https://www.ir-metadata.org/`
- **</> Google Colab notebook**
  `https://colab.research.google.com/github/irgroup/ir_metadata/blob/master/resources/demo.ipynb`
- **Dataset**
  `https://zenodo.org/record/5997491`
- **Slides**
  `https://breuert.github.io/ir-metadata-slides`
- **repro_reval**
  `https://github.com/irgroup/repro_eval`

ir-metadata.org

SIGIR 2022 Madrid

IR GROUP Technology Arts Sciences TH Köln