Search-based Recommendation: the Case for Difficult Predictions

Ghazaleh H. Torbati, Andrew Yates, Gerhard Weikum

Common Assumptions in SOTA Recommender Systems

- **Positive Only**: available datasets mostly contain positive user-item interactions.
- **All Items are Relevant**: groups of highly inter-related items exist in both training and evaluation sets of the benchmarks.
- **Context-Free Predictions**: unlike the common assumption, there is often a situative context, like a query or an example item, that should drive the prediction.
- **Uniform Negative Sampling**: in the absence of explicitly negative labels, recommenders treat all unlabeled items as negative and sample them uniformly for training and test.

Revising Assumptions for Training

- **Uniform**: the prevalent approach is to sample uniformly from all the unlabeled data.
- **Difficult**: drawing negative samples with a common category/genre to the positive training item. Idea: obtain difficult to discriminate negative training samples, w/o knowing the test-time queries.
- **Weighted**: cloning uniformly sampled unlabeled points into weighted positive and negative samples. Idea: improve learning by the more informatively labeled negative samples.

Revising Assumptions for Evaluation

- **Standard**: negative test points are drawn uniformly from all unlabeled data.
- **Profile-based**: negative test points are drawn from the category distribution of the user’s positive training set, which can be viewed as providing a user profile.
- **Search-based**: negative test points are obtained by querying all unlabeled data with the category and textual description of the positive point at hand, and keeping the 100 highest-ranked approximate matches based on BM25 retrieval scores.

System Architecture

We use a text-based approach utilizing BERT-transformer encoder to encode the textual inputs.

Evaluation Setup and Results

Amazon Books Dataset with ratings >=4. #books-per-user >= 5. #users = 1,715,645 #books = 2,066,646 #positive-interactions = 24,368,443

Conclusion

- Putting focus on the under-explored/more realistic modes of evaluation: The absolute results for the more demanding modes of evaluation are much lower.
- Proposed techniques for generating negative samples at training-time substantially improved the performance.

Contact:
ghazaleh@mpi-inf.mpg.de