The LKPY Package for Recommender Systems Experiments

Next-Generation Tools and Lessons Learned from the LensKit Project

Michael D. Ekstrand
People & Information Research Team
Boise State University
Boise, ID
michaelekstrand@boisestate.edu

ABSTRACT

Since 2010, we have built and maintained LensKit, an open-source toolkit for building, researching, and learning about recommender systems. We have successfully used the software in a wide range of recommender systems experiments, to support education in traditional classroom and online settings, and as the algorithmic backend for user-facing recommendation services in movies and books. This experience, along with community feedback, has surfaced a number of challenges with LensKit's design and environmental choices. In response to these challenges, we are developing a new set of tools that leverage the PyData stack to enable the kinds of research experiments and educational experiences that we have been able to deliver with LensKit, along with new experimental structures that the existing code makes difficult. The result is a set of research tools that should significantly increase research velocity and provide much smoother integration with other software such as Keras while maintaining the same level of reproducibility as a LensKit experiment. In this paper, we reflect on the LensKit project, particularly on our experience using it for offline evaluation experiments, and describe the next-generation LKPY tools for enabling new offline evaluations and experiments with flexible, open-ended designs and well-tested evaluation primitives.

KEYWORDS

toolkits, collaborative filtering

1 INTRODUCTION

LensKit [6] is an open-source toolkit that provides a variety of features in support of research, education, and deployment of recommender systems. It provides tools and infrastructure support for managing data and algorithm configurations, implementations of several collaborative filtering algorithms, and an evaluation suite for conducting offline experiments.

Based on our experience researching and teaching with LensKit, and the experience reports we hear directly and indirectly from others, we have come to believe that LensKit's current design and technology choices are not a good match for the current and future needs of the recommender systems research community. Further, in our examination of the software landscape to determine what existing tools might support the kinds of offline evaluation experiments we have been running with LensKit and plan to run in coming years, we came to the conclusion that there is a need for high-quality, well-tested support code for recommender systems experiments in the PyData environment.

To meet that need, we are developing LKPY, a "spiritual successor" to LensKit written in Python. This project brings the LensKit's focus on reproducible research supported by well-tested code to a more widely-used and easier-to-learn computational environment.

In this paper, we reflect on some of the successes and failures of the LensKit project and present the goals and design of the LKPY software. We invite the community to provide feedback on how this software does or does not meet their needs.

2 LENSKIT IN USE

In its 8 years of development, we and others have successfully used LensKit across all its intended application contexts.

2.1 Research

LensKit has supported a good number of research projects spanning a range of research questions and methodologies; a complete list of known published papers, theses, and dissertations using LensKit is available from the project web site[1].

We have used LensKit ourselves both for offline evaluations exploring various aspects of algorithm and user behavior [8, 9, 15] and studying the evaluation process itself [7]. It has also been used to study specialized recommendation problems, including cold start [13] and reversible machine learning [3]. Its algorithms have been used to recommend books [21], tourist destinations [22], videos [24], and a number of other item types.

One of LensKit's primary objectives is to promote reproducible, reliable research. To that end we have been modestly successful; in our own work, it has enabled us to publish complete reproducer code for recent papers [7, 9], and most recently to provide reproducer code during the review process [10].

2.2 Education

We have used LensKit to support recommender systems education in multiple settings. At the University of Minnesota, Boise State University, and Texas State University, we and our collaborators have used it as the basis for assignments in graduate classes on recommender systems. It also forms the basis for the assignments in the *Recommender Systems* MOOC [16].

[1] http://lenskit.org/research

REVEAL'18, October 7, 2018, Vancouver, BC, Canada

Response to the software in this setting has been mixed. Its standardized APIs and integration with the Java ecosystem have made things such as automated grading in the MOOC environment relatively easy to implement, but students have complained about needing to work in Java and about the system's complexity.

2.3 Production

LensKit powers the MovieLens movie recommender system [11], the BookLens book recommender [14], and the Confer conference session recommender [25]. This work, particularly in MovieLens, has enabled new user-centered research on recommender system capabilities and user response.

3 REFLECTIONS ON LENSKIT

Our direct experience using LensKit in this range of applications, and what we hear from other users and prospective users, has given us a somewhat different perspective on the software than we had when we developed the early versions of LensKit. We think some of our decisions and goals hit the mark and we were successful in achieving them; some other design decisions have not held up well in the light of experience.

3.1 What We Got Right

There are several things that we think we did well in the original LensKit software; some of these we carry forward into LKPY, while others were good ideas in LensKit's contexts but are not as important today.

3.1.1 Testing. In the LensKit development process, we have a strong focus on code testing [4]. This has served us well, and helped ensure the reliability of the LensKit code. It is by no means perfect, and there have been bugs that slipped through, but the effort we spent on testing was a wise investment.

3.1.2 The Java Platform. When we began work on LensKit, there were three possible platforms we seriously considered: Java, Python, and C++. At that time, the Python data science ecosystem was not what it is today; NumPy and SciPy existed, but Pandas did not. Pure Python does not have the performance needed for efficient recommender experiments. Java enabled us to achieve strong computational performance in a widely-taught programming language with well-established standard practices, making it significantly easier for new users (particularly students) to adapt and contribute to it than we likely would have seen with a corresponding C++ code base.

3.1.3 Modular Algorithms. LensKit was built around the idea of modular algorithms where individual components can be replaced and reconfigured. In the item-item collaborative filter, for example, users can change the similarity function, the neighborhood weighting function, input rating vector normalizations, item neighborhood normalizations, and the strategy for mapping input data to rating vectors. This configurability has been useful for exploring a range of configuration options for algorithms, at the expense of larger hyperparameter search spaces.

3.2 What Doesn't Work

Other aspects of LensKit's design and development do not seem to have met our needs or those of the community as well.

3.2.1 Opinionated Evaluation. While LensKit's algorithms were highly configurable, its offline evaluation tools are much more opinionated. You can specify a few types of data splitting strategies, and recommendation candidate strategies, and it has a range of evaluation metrics, but the overall evaluation process and methods for aggregating metric results are fixed. Metrics are also limited in their interface.

As we expanded our research into the recommender evaluation process itself, we repeatedly ran in to limits of this evaluation strategy and had to write new evaluation code in a fairly heavy framework, or just have the evaluator dump intermediate files that we would reprocess in R or Python, in order to carry out our research. Too often the answer we would have to give to questions on the mailing list or StackOverflow is "we're sorry, LesnKit can't do that yet".

One of our goals was to make it difficult to do an evaluation wrong: we wanted the defaults to embody best practices for offline evaluation. However, best practices have been sufficiently unknown and fast-moving that we now believe this approach has held our research back more than it has helped the field. It is particularly apparent that a different approach is necessary to support next-generation offline evaluation strategies, such as counterfactual evaluation [2], and carry out the evaluation research needed to advance our understanding of effective, robust, and externally valid offline evaluations.

3.2.2 Indirect Configuration. LensKit is built on the dependency injection principle, using a dependency injection container [5] to instantiate and connect recommender components. This method gave us quite a few useful capabilities, such as the automatically detecting components that could be shared between multiple experiment runs in a parameter tuning experiment, but resulted in a system where it is difficult to configure an algorithm, and difficult to understand an algorithm's configuration. It was also difficult to document how to use the system.

We believe this largely stems from the role of inversion of control in working with LensKit code — users never write code that assembles a LensKit algorithm, they simply ask LensKit to instantiate one and LensKit calls their custom components.

3.2.3 Implicit Features. Beyond indirect configuration, LensKit has a lot of implicit behavior in its algorithms and evaluator. This has at least two downsides: first, it is less clear from reading a configuration precisely what LensKit will do, making it more difficult to review code and experiment scripts; second, if documentation slipped behind the code, understanding the behavior of LensKit experiment scripts required reading the LensKit source code itself.

3.2.4 Living in an Island. LensKit has its own data structures and data access paradigms. Part of this is due to lack of standardized, high-quality scientific data tooling that is not connected to a larger framework such as Spark (while Spark does seem to expose data structures as a separate library, documentation is nonexistent).

This makes it difficult, however, to make LensKit interoperate with other software and data sets. While LensKit is flexible in the data it accepts, users must write data adapters. Using other tools such as Spark is difficult. It's unclear what, given the Java commitment, we could have done differently here, but it is definitely a liability for the future of recommender systems research.

4 TOOLKIT DESIDERATA

To address these weaknesses and power the next generation of recommender systems research, both in our own research group and elsewhere, there are several things that we desire from our recommender systems research software:

Build on Standard Tools There are now standard tools, such as Pandas and the surrounding PyData ecosystem [17], that are widely adopted for data science and machine learning research. Building on these packages maximizes interoperability with other software packages and enables code reuse across different types of research. We also find Jupyter notebooks to be a valuable means of promoting reproducibility, and would like an experiment workflow where the final analysis consists of loading the recommender results into a Jupyter notebook and computing desired metrics over them. Small experiments could even be driven entirely from Jupyter.

Leverage Existing Software Modern recommender systems are usually machine learning systems; a recommender research toolkit should not try to reinvent that wheel. Scikit-Learn provides many machine learning algorithms suitable for recommender systems research, and Keras, PyTorch, and TensorFlow all provide neural network functionality to Python. Recommender research tooling should work seamlessly with algorithms implemented in these kinds of toolkits.

Expose the Data Pipeline LensKit hides the data pipeline: it controls data splitting, algorithm training, recommendation, and evaluation. Outputs of each stage can be examined, but not manipulated, and the pipeline itself cannot be easily changed. Putting the user in control of the pipeline, and providing functions to implement standard versions of each of its stages, will have two benefits: the actual pipeline used is clearly documented in the experiment code, and the pipeline can be modified as research needs demand.

Explicit is Better than Implicit Related to exposing the data pipeline, we wish for our new tools to follow the Python philosophy of favoring explicit denotations of desired operation. This will make it easier to review experiment designs and improve the reliability and rigor of reproducible research. We hope that this will also make the LKPY code itself easier to read and understand.

Simple Interfaces Interfaces to individual software components should be as simple as possible, so that it is easy to document, test, and reimplement them.

Easy-to-Use Development Environment In order for prospective users and contributors, particularly students, to use LKPY and participate in its development, we want the development tools to be as standard and easy-to-use as possible.

Notably absent from this list is LensKit's (in)famous algorithm configurability. That configurability was useful for exploring the space of algorithm configurations, but its particular design is more suitable to heuristic techniques such as k-NN; machine learning approaches seem better served by a different design. Connecting with existing flexible optimization software will provide a great deal of configurability for new algorithms. We think it is more important for the recommender-specific software to focus on flexibility in the experiment design.

5 THE LKPY SOFTWARE PACKAGE

To that end, we are developing a successor to LensKit, LKPY. LKPY is a new Python package for recommender systems experiments, particularly offline evaluations and similar studies, that we hope will also be useful in educational settings. The first version of LKPY is now available[2], and is capable of running many of the kinds of experiments we have run with LensKit in a more flexible and forward-looking fashion.

LKPY is available under the MIT license.

5.1 LKPY Facilities

LKPY provides several modules for aiding recommender experiments:

Data Preparation The `crossfold` module provides functions for splitting data for cross-validation. It supports rating-based, user-based, and item-based splitting strategies, with configurable data holdout settings.

Algorithm APIs The `algorithms` module defines Python interfaces for training models, generating predictions, and generating recommendations. These interfaces are minimal, defined in terms of Pandas data structures, and can be readily implemented on top of any desired machine learning toolkit such as Scikit-Learn or Pandas.

Evaluation Metrics The `metrics` package provides classical top-N and prediction accuracy metrics. Metrics are functions that operate directly over Pandas series objects, and thus can be applied to any data of an appropriate shape. There is therefore no limit to how the user reprocesses recommendations and predictions before computing metrics.

Classical CF Algorithms LKPY provides implementations of nonpersonalized, k-NN, and biased matrix factorization collaborative filters. We expect to expand the set of algorithms provided, but being a source of algorithms is not LKPY's primary objective. These algorithms are provided in part to give LensKit users a migration path to LKPY that keeps the algorithms as consistent as possible; LKPY's algorithm implementations are based on LensKit's and should generally be the same. They are less configurable, however, selecting configuration options such as item-item similarity functions that we have found to work well across a range of data sets. We use Cython to accelerate inner algorithm computations when NumPy, SciPy, or Pandas operations are inadequate.

Batch Utilities To ease writing evaluation scripts, LKPY provides utility functions for computing predictions or recommendations for many users in batch.

[2] https://lkpy.lenskit.org

```
import pandas as pd
from lenskit import batch, topn
from lenskit import crossfold as xf
from lenskit.algorithms import knn

ratings = pd.read_csv('ml-100k/u.data', sep='\t',
        names=['user', 'item', 'rating', 'timestamp'])

algo = knn.ItemItem(30)

def eval(train, test):
    model = algo.train(train)
    users = test.user.unique()
    recs = batch.recommend(algo, model, users, 100,
            topn.UnratedCandidates(train))
    # combine with test ratings for relevance data
    res = pd.merge(recs, test, how='left',
                   on=('user', 'item'))
    # fill in missing 0s
    res.loc[res.rating.isna(), 'rating'] = 0
    return res

# compute evaluation
splits = xf.partition_users(ratings, 5,
        xf.SampleFrac(0.2)
recs = pd.concat((eval(train, test)
                  for (train, test) in splits))

# compile results
ndcg = recs.groupby('user').rating.apply(topn.ndcg)
```

Figure 1: Example Simple Evaluation

Since components connect with standard Pandas data structures, they can be used together or individually. It is trivial to use alternative algorithms with LKPY's data splitting and metrics, or to use an entirely different data preparation strategy with LensKit's algorithms and batch functions.

One way in which our desire to favor explicit code over implicit behavior is in data processing: instead of telling LensKit how to transform data, in LKPY, the user just directly writes their data transformations using standard Pandas operations.

5.2 LKPY Example Code

Figure 1 shows an example of using LKPY to compute the nDCG of a k-NN algorithm with 5-fold cross-validation on the MovieLens 100K data set. It is somewhat verbose, but ever step of the evaluation process is clear in the code, so the experiment structure can be checked and it is self-documenting when the evaluation code is published.

5.3 Dependencies and Environment

LKPY leverages Pandas [18], numpy [19], scipy [20], and PyTables/HDF5, along with several other Python modules. We use Cython [1] for native-code acceleration, as it integrates with numpy and OpenMP with a minimum of boilerplate.

We regularly test LKPY with recent Python versions on Windows, Linux, and macOS. Our ongoing criteria are to support the three major operating systems, and the version of Python 3 available in the most most recent release of major Linux distributions (Ubuntu LTS, RHEL/CentOS with EPEL, and Debian). We focus primarily on Anaconda-based Python installations, but also test the code with vanilla Python on all supported platforms. We provide binaries via Anaconda Cloud for supported Python versions and operating systems[3], along with source distributions on the Python Package Index[4]. We are evaluating providing precompiled binaries through the Python Package Index as well.

5.4 Future Directions

We intend to fill out some more algorithms in LKPY's capabilities, such as out-of-the-box integrations for SLIM, BPR, and Poisson factorization, and add facilities for alternative algorithm evaluation strategies. We also plan to build bridges to enable the algorithm implementations provided by other packages such as surprise [12] and PyRecLab [23]. LKPY provides a clean, minimal core that can be extended however our needs and those of the research community require.

6 COMPARISON TO EXISTING PACKAGES

Before embarking on this project, we re-examined the landscape of existing software to determine if the needs we saw would be met by one of the other software packages. We were unable to find existing tooling that supports our goals of flexible recommender systems experiments that leverage the PyData ecosystem.

Two of the most obvious contenders for current Python recommender systems are surprise [12] and PyRecLab [23]. While Surprise uses numpy and scipy, neither of these packages leverage the PyData ecosystem; instead, each has its own classes for representing data sets. PyRecLab's evaluation facilities are also more automatic; we want to write out the evaluation steps in our own code so that we can experiment with them more readily. Further, we do not want to require students to learn C++ to be able to particpate in research that requires extending LKPY.

Our approach is perhaps most similar to that of mrec[5], in that we focus on discrete steps enabled by separate tools. Again, though, mrec does not leverage contemporary Python data science tools, and does not seem to be under active maintenance. We also focus on Python as the scripting language for experiment control instead of mrec's command-line orientation.

It is our sense that many Python-based recommender systems research projects roll their own evaluation procedure directly in Python tools while building the recommender in scikit-learn or one of the deep learning frameworks. It is our goal to integrate with such workflows, enabling them to leverage common, well-tested implementations of metrics and other experimental support code while continuing to use their existing data flows for the recommendation process.

[3] https://anaconda.org/lenskit/lenskit
[4] https://pypi.org/project/lenskit/
[5] https://github.com/Mendeley/mrec

7 CONCLUSION

We have learned many lessons developing, maintaining, and researching with the LensKit software. Based on these lessons, we came to the conclusion that, in its current form, it is not meeting our needs or the needs of the recommender systems community well, and that resources would be better spent on tools that improve research being done with the widely-used Python packages that drive much of modern data science.

Where LensKit focused on providing building blocks for recommender *algorithms*, LKPY provides building blocks for recommender *experiments*. Built on the PyData stack and oragnized around clear, explicit data processing pipelines with a minimum of custom concepts, LKPY provides a solid foundation for new experimentation and concepts at all stages of the offline recommender system evaluation lifecycle. We expect it to meet our own research needs in offline evaluation and simulation of recommender systems much better than the current LensKit code going forward, and hope that others in the research community find it useful as well. We welcome contributions on GitHub, and invite feedback from the community to guide future development.

REFERENCES

[1] S Behnel, R Bradshaw, C Citro, L Dalcin, D S Seljebotn, and K Smith. 2011. Cython: The Best of Both Worlds. *Computing in Science Engineering* 13, 2 (March 2011), 31–39. https://doi.org/10.1109/MCSE.2010.118

[2] LÃĺon Bottou, Jonas Peters, Joaquin QuiÃśnero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of machine learning research: JMLR* 14, 1 (2013), 3207–3260. http://www.jmlr.org/papers/volume14/bottou13a/bottou13a.pdf

[3] Yinzhi Cao and Junfeng Yang. 2015. Towards Making Systems Forget with Machine Unlearning. In *Proceedings of the 36th IEEE Symposium on Security and Privacy*. IEEE. http://www.ieee-security.org/TC/SP2015/papers-archived/6949a463.pdf

[4] Michael Ekstrand. 2016. Testing Recommenders. https://buildingrecommenders.wordpress.com/2016/02/04/testing-recommenders/ Accessed: 2017-1-6.

[5] Michael D Ekstrand and Michael Ludwig. 2016. Dependency Injection with Static Analysis and Context-Aware Policy. *Journal of Object Technology* 15, 1 (Feb. 2016), 1:1. https://doi.org/10.5381/jot.2016.15.5.a1

[6] Michael D Ekstrand, Michael Ludwig, Joseph A Konstan, and John T Riedl. 2011. Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and LensKit. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, 133–140. https://doi.org/10.1145/2043932.2043958

[7] Michael D Ekstrand and Vaibhav Mahant. 2017. Sturgeon and the Cool Kids: Problems with Top-N Recommender Evaluation. In *Proceedings of the 30th Florida Artificial Intelligence Research Society Conference*. AAAI Press. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/viewPaper/15534

[8] Michael D Ekstrand and John T Riedl. 2012. When recommenders fail: predicting recommender failure for algorithm selection and combination. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 233âĂŞ236. https://doi.org/10.1145/2365952.2366002

[9] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, Pera, and Maria Soledad. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (PMLR)*, Vol. 81. 172âĂŞ186. http://proceedings.mlr.press/v81/ekstrand18b.html

[10] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the Twelfth ACM Conference on Recommender Systems*. ACM.

[11] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (Dec. 2015), 19:1âĂŞ19:19. https://doi.org/10.1145/2827872

[12] Nicolas Hug. 2017. Surprise, a Python library for recommender systems. http://surpriselib.com.

[13] Daniel Kluver and Joseph A Konstan. 2014. Evaluating Recommender Behavior for New Users. In *Proceedings of the Eighth ACM Conference on Recommender Systems (RecSys '14)*. ACM. https://doi.org/10.1145/2645710.2645742

[14] Daniel Kluver, Michael Ludwig, Richard T. Davies, Joseph A. Konstan, and John T. Riedl. 2014. BookLens. https://booklens.umn.edu/

[15] Daniel Kluver, Tien T Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How many bits per rating?. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 99âĂŞ106. https://doi.org/10.1145/2365952.2365974

[16] Joseph A Konstan, J D Walker, D Christopher Brooks, Keith Brown, and Michael D Ekstrand. 2015. Teaching Recommender Systems at Large Scale: Evaluation and Lessons Learned from a Hybrid MOOC. *ACM Transactions on Computer-Human Interaction* 22, 2 (April 2015), 10:1âĂŞ10:23. https://doi.org/10.1145/2728171

[17] Wes McKinney. 2018. *Python for Data Analysis: Data Wrangling with pandas, NumPy, and IPython*. O'Reilly. http://shop.oreilly.com/product/0636920023784.do

[18] Wes McKinney and Others. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. 51–56. http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

[19] Travis E Oliphant. 2006. *A Guide to NumPy*. Trelgol Publishing.

[20] T E Oliphant. 2007. Python for Scientific Computing. *Computing in Science Engineering* 9, 3 (May 2007), 10–20. https://doi.org/10.1109/MCSE.2007.58

[21] Maria Soledad Pera and Yiu-Kai Ng. 2017. Recommending books to be exchanged online in the absence of wish lists. *Journal of the Association for Information Science and Technology* (Nov. 2017). https://doi.org/10.1002/asi.23978

[22] Toon De Pessemier, Jeroen Dhondt, and Luc Martens. 2016. Hybrid group recommendations for a travel service. *Multimedia tools and applications* 75, 5 (Jan. 2016), 1–25. https://doi.org/10.1007/s11042-016-3265-x

[23] Gabriel Sepulveda and Denis Parra. 2017. pyRecLab: A Software Library for Quick Prototyping of Recommender Systems. *arXiv preprint arXiv:1706. 06291* (2017). https://arxiv.org/abs/1706.06291

[24] Marius LÃżrstad Solvang. 2017. *Video Recommendation Systems: Finding a Suitable Recommendation Approach for an Application Without Sufficient Data*. Master's thesis. http://hdl.handle.net/10852/59239

[25] Amy X. Zhang, Anant Bhardwaj, and David Karger. 2016. Confer: A Conference Recommendation and Meetup Tool. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*. ACM, New York, NY, USA, 118–121. https://doi.org/10.1145/2818052.2874340