Sparse Approximate Inverse for Enhanced Scalability in Recommender Systems

FACULTY OF MATHEMATICS AND PHYSICS Charles University

Mgr. Martin Spišák | Supervisor: Mgr. Ladislav Peška, Ph.D.

Charles University, Faculty of Mathematics and Physics. In collaboration with GLAMI.

GLAMI

Open-source model

Motivation

Shallow neural networks are simple yet often outperform deep learning approaches in collaborative filtering tasks [1]. **Embarrassingly Shallow Autoencoder** ($EASE^R$) [2] is a linear model, which – despite its simplicity –

- aggregates feedback from all users to compensate for scarce feedback from individuals
- uses *long chains* of user–item feedback to model item similarity.

Instead of gradient descent, the training procedure uses *closed-form solution* of its convex optimization objective, improving training complexity. However, this process relies on the calculation of $A^{-1} = (X^\wedge TX + \lambda I)^{-1}$, introducing **two challenges for practical application:**

1. Computing A^{-1} is **costly** (but depends only on #items).
2. Despite the sparsity of input data X, A^{-1} (and also the weights) will be **dense**.

Crucially, the model must fit in RAM for inference.
1M items → model size = 4 TB (in `float32`).

Conclusion

Popular shallow autoencoder $EASE^R$ leverages long user–item interaction chains. This ability positively affects the quality of recommendations but also prohibitively increases training and inference costs on large item sets. We introduce a solution to these problems using modern numerical methods for sparse approximate inversion. The techniques are scalable and robust enough to find critical (even long-distance) information. By exploiting the inherent sparsity of user–item interaction data, our end-to-end sparse method achieves substantial efficiency gains over previous approaches that attempt to overpower the problem using dense block operations. The resulting model SANSA provides a robust yet attainable baseline model for researchers with limited resources and large-scale industry environments with millions of items.

The thesis outcomes were presented at an international conference on recommender systems [3]. The model is currently under testing for production deployment.

Highlights of the proposed method

- Alleviate the main drawback of a broadly used $EASE^R$ recommendation algorithm
- **Cheap & easy-to-use** for researchers & **scalable** enough even for large industrial settings

Experiment results

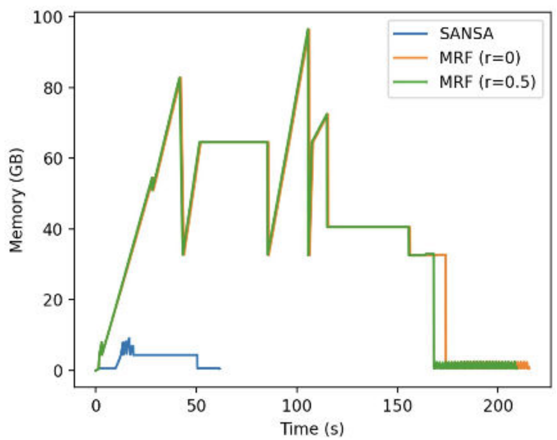
- demonstrate robustness and efficiency on 5 datasets
- Amazon Books: 53K users, **92K items**, 3M interactions

Amazon Books

	SANSA (ICF)	MRF ($r=0$)	MRF ($r=0.5$)	results reprinted from [6]: $EASE^R$	SLIM	ITEMCF	ULTRAGCN
recall@20	**0.077**	0.071	0.069	0.071	0.075	0.074	0.068
nDCG@20	**0.064**	0.058	0.055	0.057	0.060	0.061	0.056
Training resources							
vCPU	**2**	16	16	28	28	28	20*
memory usage (GB):							
peak	**9.18**	96.45	96.58	—— not measured; costly ——			
average	**3.87**	49.12	49.75	—— not measured; costly ——			
time	**49 s**	172 s	167 s	222 m	316 m	57 m	45 m

*and a GPU (RTX 2080)

- 3x faster training with 10x less memory compared to previous sparse modification of $EASE^R$ – MRF [4]
- **orders of magnitude faster and cheaper** than other models
- new state-of-the-art accuracy on the dataset



How to scale $EASE^R$ to *millions* of items?

Approximate $EASE^R$ using a sparse model

- preserve properties of A^{-1} --- full rank, SPD
- enable **arbitrary model compression** --- allow users to specify weight density of the resulting model

Method: **factorized sparse approximate inversion**

- sophisticated approaches developed for numerical solvers [5]
- **extract global dominant information** from user--item interaction graph
- A is SPD → increased efficiency, higher compression

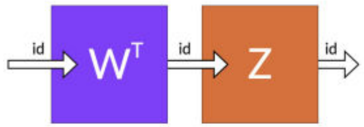
The approximate inverse is computed in 3 steps:

1. approximate (or incomplete) *sparse* Cholesky factorization
2. *free* initial approximation of the inverse factor
3. refinement based on Frobenius norm minimization

Model – training and architecture

Scalable Approximate NonSymmetric Autoencoder (SANSA)

input user–item interaction matrix X, L2 regularization λ
compute sparse $LDL^T \approx P(X^TX + \lambda I)P^T$ (for a permutation P)
compute sparse $K \approx L^{-1}$
$W \leftarrow KP$
$Z_0 \leftarrow D^{-1}W$
$\vec{r} \leftarrow \mathrm{diag}(W^TZ_0)$
$Z \leftarrow$ scale the columns of Z_0 by $-1/\vec{r}$
return W^T, Z



References

[1] Yushun Dong, Jundong Li, and Tobias Schnabel. When newer is not better: Does deep learning really benefit recommendation from implicit feedback? In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, page 942–952, New York, NY, USA, 2023. Association for Computing Machinery.
[2] Harald Steck. Embarrassingly shallow autoencoders for sparse data. In The World Wide Web Conference, WWW '19, page 3251–3257, New York, NY, USA, 2019. Association for Computing Machinery.
[3] Martin Spišák, Radek Bartyzal, Antonín Hoskovec, Ladislav Peška, and Miroslav Tůma. 2023. Scalable Approximate NonSymmetric Autoencoder for Collaborative Filtering. In Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604915.3608827
[4] Harald Steck. Markov random fields for collaborative filtering. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019. Curran Associates Inc.
[5] M. Benzi and M. Tůma. A comparative study of sparse approximate inverse preconditioners. ANM, 30(2-3):305–340, 1999.
[6] The BARS Community. Barsmatch: A benchmark for candidate item matching, 2023.