Sparse Approximate Inverse for Enhanced Scalability in Recommender Systems

Mgr. Martin Spišák | Supervisor: Mgr. Ladislav Peška, Ph.D.
Charles University, Faculty of Mathematics and Physics. In collaboration with GLAMI.

Motivation

Shallow neural networks are simple yet often outperform deep learning approaches in collaborative filtering tasks [1]. Embarrassingly Shallow Autoencoder (EASE©) [2] is a linear model, which – despite its simplicity – aggregates feedback from all users to compensate for scarce feedback from individuals. It 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 $A^{-1} = (X^TX + \lambda I)^{-1}$, introducing two challenges for practical application:

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

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

Conclusion

Popular shallow autoencoder EASE© leverages long user-item interaction chains. This ability positively affects the quality of recommendations but also proportionally 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© 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

<table>
<thead>
<tr>
<th></th>
<th>SANSA (ICF)</th>
<th>MRF (r = 0)</th>
<th>MRF (r = 0.5)</th>
<th>EASE©</th>
<th>SLIM</th>
<th>ITEMCF</th>
<th>ULTRAGCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>recall@20</td>
<td>0.077</td>
<td>0.071</td>
<td>0.069</td>
<td>0.071</td>
<td>0.075</td>
<td>0.074</td>
<td>0.068</td>
</tr>
<tr>
<td>nDCG@20</td>
<td>0.064</td>
<td>0.058</td>
<td>0.055</td>
<td>0.057</td>
<td>0.060</td>
<td>0.061</td>
<td>0.056</td>
</tr>
<tr>
<td>vCPU</td>
<td>2</td>
<td>16</td>
<td>16</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>20*</td>
</tr>
<tr>
<td>memory usage (GB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>peak</td>
<td>9.18</td>
<td>96.45</td>
<td>96.58</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>average</td>
<td>3.67</td>
<td>49.12</td>
<td>49.42</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>time</td>
<td>49 s</td>
<td>172 s</td>
<td>167 s</td>
<td>222 m</td>
<td>316 m</td>
<td>57 m</td>
<td>45 m</td>
</tr>
</tbody>
</table>

3x faster training with 10x less memory compared to previous sparse modification of EASE© – MRF [4]

orders of magnitude faster and cheaper than other models

new state-of-the-art accuracy on the dataset

How to scale EASE© to millions of items?

Approximate EASE© 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 $\rightarrow$ 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)

1. input user–item interaction matrix $X$, L2 regularization $\lambda$
2. compute sparse LDLT $\approx P/(X'X + \lambda I)^{1/2}$ (for a permutation $P$)
3. compute sparse $K = L^{-1}
4. $W \leftarrow KP$
5. $Z_0 \leftarrow D^{-1}W$
6. $\tilde{Z} \leftarrow \text{diag}(W'Z_0)$
7. $Z \leftarrow \text{scale the columns of } Z_0 \text{ by } 1/\tilde{Z}$
8. return $W$, $Z$

References