Skoltech

Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Evgeny Frolov

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Low-rank models for recommender systems with limited preference information

Supervisor: Prof. Ivan Oseledets

Chair of PhD defense Jury: Prof. Andrzej Cichocki **Email:** *a.cichocki@skoltech.ru*

Date of Thesis Defense: 19 September 2018

Name of the Reviewer: István Pilászy (PhD)

I confirm the absence of any conflict of interest	**Signature:** Pilászy István **Date: 29-08-2018**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The topic of Evgeny Frolov's dissertation is finding new recommender methods that can handle the cases when only very limited information is available about users or items. This topic is important, since in online environments it is crucial to provide the users with relevant recommendations from the very first moment they start using a service.

This work focuses on enhancing matrix factorization methods, as these methods exhibit nice properties, like being able to handle very large datasets, and uncover indirect relationship between items that are not directly observable from co-occurrence patterns.

The dissertation proposes three new methods for handling these problems.

Chapter 1, 2 and 3 provides a very comprehensive survey of the literature. Chapter 1 provides an introduction to the general concepts, Chapter 2 summarizes matrix factorization methods, and Chapter 3 is about tensor factorization methods.

Chapter 4 provides a clear description about the motivation behind this dissertation. It distinguishes two important problems: local lack of preferences (when there is not enough information about a particular user or item) and global lack of preferences (when the user-item matrix is overly sparse).

Section 4.4 sets the goals, namely to create new methods that satisfies the following criteria:

- *"uses SVD or applies it sequentially for optimization,*
- *requires minimal tuning of hyper-parameters,*
- *supports highly dynamic online settings via folding-in,*
- *efficiently handles large number of different side features,*
- *provides a unified solution to the limited preference information problem and*
- *can be easily adapted to each of its subproblems."*

Chapter 5 proposes a simple, yet brilliant idea: instead of arranging rating values into the user-item matrix, let us use a third-order tensor, where the third dimension is used to encode the different values of ratings, and Tucker Decomposition is applied on that tensor. This new method is capable of providing meaningful recommendations even if the user has provided only one rating (either a positive or a negative rating), in which case other factorization methods fall short. The author also proposes new evaluation metrics that can measure the tendency of recommending irrelevant items.

Chapter 6 proposes a new method for handling side information (e.g. movie metadata) when the user-item matrix is very sparse. The proposed method is a modification of the PureSVD approach, where the user and item feature matrices are orthonormal not in the original space but in a transformed space, where the transformation depends on the side information in an unsupervised way. The side information is represented as an item-item similarity matrix. The author proposes an efficient trick here: instead of computing the square root of the item-item similarity matrix, he computes Cholesky decomposition, which is more efficient when the item-item similarity matrix is sparse. The author evaluates the proposed method both in warm-start and cold-start scenarios.

Chapter 7 unifies the proposed methods of Chapters 5 and 6. It shows how the HOOI algorithm used for calculating the Tucker Decomposition can be modified to efficiently handle the side information in the same way as it is handled in Chapter 6, in order to boost the ROC- / nDCG- / nDCL- based metrics used in Chapter 5.

Chapter 8 describes the open source framework used for the experiments, developed by the author.

The proposed 3 new methods are present in Chapters 5, 6 and 7.

The list of issues to be addressed before/during the thesis defense

On page two, references are wrong, e.g. "Resnick, Hill, Shardanand and Maes [96]", while [96] does not belong to these authors. One more example of this mistake on page two: "GroupLens [23]".

In Section 3.2, there are many methods described in detail. However, these methods are then subject to neither comparison with the proposed methods, nor enhancement, therefore it is not necessary to provide such a detailed description of tensor factorization methods.

In Section 5.4.3, you set the number of factors to 10, and CoFFee multilinear rank to (13,10,2). Please show results with larger number of factors, e.g. 100 or 1000.

In Figure 6.1, how do the results depend on the number of latent factors?

Section 6.2.1 is hard to follow. Please describe how Generalized SVD works. Please also describe how U^ can be obtained from U, and how the proposed method differs from Generalized SVD. Please also provide an Algorithm description (pseudocode for this entire proposed method).

In Section 6.2.2, please provide an example of creating such a similarity matrix. What is the sparsity of such a matrix in the datasets used for the experiments? Does using Euclidean distance or Jaccard Index instead of cosine similarity have a positive or negative impact on the final nDCG scores? Is side-information following a Zipf-like distribution? If you construct a similarity matrix using cosine similarity between some already sparse feature vectors reflecting side information, then you already have a sparse decomposition of the similarity matrix. Could this be used in place of the Cholesky decomposition?

In Section 6.2.3 you mention incomplete Cholesky decomposition. Have you tried to experiment with this?

In eq (6.6) what is the matrix A ?

After eq (6.6), you state that eq (6.6) also provides a solution to eq (6.3) with U^=... and V^=... -- this statement is hard to follow, it is really not clear what variable should be substituted in which expression.

In Section 6.3.2: Book-Crossing is not published by GroupLens.

In Section 6.3.2: please elaborate, how you construct the side similarity matrices. What is the sparsity of the obtained matrices? For Movielens-1M dataset: 41% of the movies have "Drama" as one of their genres, and for MovieLens-10M dataset, 50% of the movies have "Drama" as one of their genres. Won't this make the similarity matrix dense?

On Figure 6.3, I miss HybridSVD with alpha=0.5 from the bottom left graph, and HybridSVD with alpha=0.1 from the bottom right graph.

For some computations, you need the inverse of the Cholesky decomposition, i.e. L_S^{-T}. Is this matrix also sparse? If not, it is a problem?

Regarding the datasets / algorithms used in Section 6: although the author put a huge effort in fine-tuning Factorization Machines, I would like to see some experiments with other algorithms, or some experiments with other datasets. Both BookCrossing and MovieLens are really old datasets, there are some other really sparse datasets, for example, the Yahoo! Music dataset or the Million Song DataSet, etc. On the other hand, there are some other methods capable of handling metadata, to name a few:

- the article "Learning Attribute-to-Feature Mappings for Cold-Start Recommendations" by Gantner et al.
- the method of [145] can be easily applied on implicit domains with the help of [94]
- Field-aware Factorization Machines
- simply concatenating the side information matrix (metadata - item matrix) and the user - item matrix provides a way for regularizing the latent item features.

For Section 6, It would be interesting to see a comparison of HybridSVD with some other matrix factorization method, like WRMF (not using side info), just like it is done for Section 5.

Section 7: it would be great if the HOOI algorithm were described with pseudocode.

Figure 7.1: some graphs overlap (e.g. I can see only 4 curves on the graph located in the middle-center), it would be great if you can show these results in a tabular form as well. In the top-n evaluation scenarios I am more interested in the lower n values, but the graphs are very crowded for such n values, a tabular form would be great in this case as well.

In Table 6.2, you performed a sparsity test on the ML10M dataset. Would it be possible to repeat this experiment for Chapter 7 (Figure 7.1) as well? I mean: how would curves of Figure 7.1 look like for the ML10M dataset if it were more sparse?

Bibliography:

- for [17], there are some strange characters.
- sometimes the authors are put first, sometimes the title. For example, [46] vs [47]. Bibliography should be formatted with a coherent style.

I observed only a very few typos throughout the document:

- p8: measures → measure
- p37: he or he
- p42: {r_i}j
- p58: Sections Sec. 3.2.2 and Sec. 3.2.4 -- word repetition: "Sections" and "Sec."
- p67: Sec. 3.2.4 and Sec. 3.2.4
- p103: “ranking OR irrelevant”: OR should be OF instead.
- p125: than fixed → then fixed
- p138: CoFFee is written with uppercase F, but HybridCoffee is written with lowercase F.
- p139: in of irrelevant

What I liked most in the dissertation

Each of the proposed methods aims to be simple and practically applicable. Chapter 5 (CoFFee) is a brilliant idea to recommend for users with only one or few ratings, even if those ratings are negative.

Chapter 6 (HybridSVD) proposes a new, efficient and elegant way to handle side information with SVD, and applies a clever trick (Cholesky decomposition) to avoid the computation of the square root of the similarity matrix.

Chapter 7 (HybridCoFFee) unifies the previous two chapters, thus these 3 chapters form a complete work.

Table 6.2: It is really interesting to see that PureSVD can go below most popular recommendations, when the matrix is very sparse, but HybridSVD can help in this situation.

The relevance of the methods used in the dissertation

The proposed methods are relevant, matrix factorization is still an actively researched area in the field of recommender systems, although deep learning and treating users as a sequence of events are two emerging directions.

The scientific significance of the results obtained and their compliance with the international level and current state of the art

The proposed methods provide significant performance improvement over baseline methods. I would like to see some additional experiments with more datasets / algorithms.

The CoFFee algorithm tackles a very important but neglected area of recommender systems, namely providing meaningful recommendations for users with few or negative feedback. As the author shows, other methods perform poorly on this task. From a practical viewpoint, this task is way more important than recommending for users with plenty of feedback. The HybridSVD method provides a new way to handle the cold start, and HybridCoFFee unifies these two algorithms.

The relevance of the obtained results to applications (if applicable)

Providing relevant recommendations to users from the very first time is a really important task. Being able to handle really sparse user-item matrices is also really important. In practice, this is the main case, and dense user-item matrices are the exceptions.

The proposed methods excel in both cases, which makes the entire dissertation a really valuable work.

The quality of publications

The author has 5 publications. In all of his publications, he is the first author. Two publications provide a survey of the recommender system literature and demonstrates that the author has a deep knowledge of recommender systems. The 3 other publications are the basis for the 3 proposed new methods. The method of Chapter 5 is published on RecSys, the most important conference of recommender systems, while the methods of Chapter 6 and Chapter 7 are published only on arXiv.

Questions for the author

1. Is it possible to combine PureSVD and HybridSVD, by using HybridSVD only for less popular items?

2. Figure 7.1: for the BX dataset, CoFFee does not perform well for nDCG and nDCL. What could be the reason?

3. For HybridSVD and PureSVD, do you handle user or item bias?

4. For HybridCoFFee: would it be possible to extend this method for 4 dimensional tensors?

5. What are the possible research directions to enhance the proposed algorithms further?

6. What are the most important open questions of recommender systems?

7. What is the future of matrix and tensor factorization methods in the field of recommender systems?

Provisional Recommendation

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*