Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

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: September 19, 2018

The thesis document includes the following changes in answer to the external review process.

Reviewer: Marko Tkalcic

Reviewer Comment 1
In chapter 1 he provides an overview of the field and exposes the following challenges: cold start, missing values, usage of implicit feedback, evaluation and reproducibility aspects, real-time issues, context and features. I would recommend to conclude this chapter with specifying which of these challenges will be addressed in the remainder of the thesis.
Answer
Thank you for bringing our attention to it. There's now a new section in chapter 1 called "*Quick summary and outlook*", which links the described challenges to concrete aspects of the developed methods and provides the necessary forward references to the corresponding chapters.
Reviewer Comment 2
The formatting of the bibliography is inconsistent. Some entries start with the title and some with the authors.
Answer
Thank you for noting this. Fixed.
Reviewer Comment 3
In chapter 2 the candidate provides a thorough overview of factorization techniques, their advantages and limitations. Similarly, in chapter 3, the candidate gives an overview of tensor factorization techniques. I would recommend that each of these chapters concludes with the concrete downside that the thesis is addressing. Furthermore, for each research problem identified, the author should point to the relevant chapter (5/6/7) that addresses it.
Answer
Thank you for this suggestion. Both chapter 2 and chapter 3 now contain a summary of problems that are to be addressed in the thesis with the necessary forward links to the corresponding chapters 4/5/6/7.
Reviewer Comment 4
the conclusion of chapter 3 is "A possible cure for this problem is to use TT/HT decomposition. In our opinion, this is a promising direction for further investigations." It is not entirely clear whether the author addresses this problem with the proposed method later on in the thesis or it is a proposal for future work based on an educated guess.

Answer

Thank you for pointing this out. This particular statement was moved to the section with final conclusions as a future research direction.

Reviewer Comment 5

In chapter 4, the candidate describes the problem of missing data, referred to as limited preference information. The author distinguishes between the local and global lack of preferences. However, the distinction is not very clear. Is the global a matter of global sparsity and local only user- or item-bound? Please, rephrase to make it clearer.

Answer

We agree. Introduction to chapter 4 was rewrtitten in order to draw a clearer distinction between the two types of problems from the very beginning without requiring the reader to wait until later sections. The following intuition is provided: global problem is driven by the very mechanism of interactions, it is permanent and affects the system at any time. The local problem is considered to be temporary and tends to disappear with time, when more data is collected.

Reviewer Comment 6

The candidate argues that there is little related work that uses negative ratings. While this is true to a certain extent, there are techniques for eliciting negative feedback, such as the MinRating. I would recommend comparing active learning techniques for preference elicitation, e.g. M Elahi, F Ricci, N Rubens. A survey of active learning in collaborative filtering recommender systems. - Computer Science Review, 2016.

Answer

Thank you for providing the related references! Section 4.3.1 has been extended accordingly. The reference to an overview of active learning approaches is included. The text now also includes additional discussion of MinRating and related techniques and shows their connection to the problems considered in the thesis.

Reviewer Comment 7

In Sec. 4.4 the candidate lays out the requirements for improvements over state-of-the-art methods, described in Sects 4.3.1 and 4.3.2. However, in these two sections, the author has briefly described the existing methods, but has failed to clearly identify the downsides of these methods. For example, stating that those methods do not use SVD-based techniques is not a valid argument for introducing SVD. Epy remainder of the section 4.4 is similarly unconvincing. For example, the author states "Moreover, these methods focus on a particular subproblem." Why is this a downside? Please elaborate more on the downsides of existing work in order to justify and strengthen your scientific contributions.

Answer

Thank you for emphasizing this. Both sections now contain more detailed description of the disadvantages of existing methods with additional examples and more thorough explanation of the benefits of the proposed methods. We also agree that the statement about solving particular subproblems does not provide enough context to make a concrete point. We removed this phrase and used different wording to note that we aim at creating a single approach, which is applicable in both local and global problem settings.

Reviewer Comment 8

In chapter 5 … He used the Movielens dataset and a couple of baseline algorithms. It is not clear why there is no FM among these as at the beginning of the chapter he states "This type of relations can be modelled with several methods, such as Factorization Machines [67] or other contextaware methods [42]."

Answer

We agree that having FM in the list of baseline algorithms would make the work more solid. On the other hand, technically, it would not be the classic FM implementation. Most importantly, as we operate in the warm start regime, i.e., on previously unseen users, it would require to implement our own folding-in procedure, as it is not provided out-of-the-box by popular FM libraries. There are, indeed, some works on incremental approach for FM algorithm, however it was never proposed

for our particular problem of a more appropriate treatment of ratings, which has its own specifics and would essentially require implementing and testing several ideas from scratch. For this reason it was decided not to go with FM in this particular part of the work.

Reviewer Comment 9

There are details missing on how baselines were implemented. E.g. how were negative ratings used in BPR and WRMF? How were Movielens data converted to pairwise (for BPR)?

Answer

Thanks for pointing this out. We used standard MyMediaLite, which accepts positive-only data. We had to cut-off ratings below the positivity threshold value and the remaining part was fed into the library, as it requires. The data was further handled solely by the library. Section 5.4.3 is extended with the explanations accordingly.

Reviewer Comment 10

In chapter 6, the candidate presents a hybrid model. He starts with the statement "To the best of our knowledge there were no attempts to build a hybrid SVD-based approach where interaction data and side information would be factorized jointly in a seamless way". Could you elaborate a bit more? Especially what do you mean by "seamless"? There are factorization approaches where additional data is injected as additional (not really latent) factors, e.g. Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., & Cantador, I. (2016). Alleviating the new user problem in collaborative filtering by exploiting personality information. User Modeling and User-Adapted Interaction, 26(2), 1–35. https://doi.org/10.1007/s11257016-9172-z.

Answer

We agree that this is probably too vague definition. We have replaced the sentence with a more explicit description and provided example to contrast our approach to the other approaches, where SVD is used as an intermediate tool, rather than an end model. The key idea that we try to convey, is that it is possible to stay within the computational paradigm of the classical SVD algorithm and to not turn to other optimization algorithms, which would be inevitable in the case of a more flexible problem formulation, like the one that you are referring to. Staying with SVD, in turn, has many practical advantages, which we provide in chapter 2 and summarize in chapter 4. While our proposed method is not that flexible, we believe it still can find its applications in certain scenarios.

Reviewer Comment 11

I would recommend to introduce an example of side information in order to make figure 6.2 more understandable. At this point it is still not clear which side information is used and why that behavior is observed.

Answer

We have added a simple explanation of how matrix S can actually be computed in the beginning, where it is introduced for the first time. We also provide some practical example, which should help to resolve possible confusion and help further reading.

Reviewer Comment 12

I would recommend to report also the computation times and adding a discussion on the trade-off between performance and accuracy.

Answer

We absolutely agree that actual computational performance is an important part of the experiment. We report computation times in Sec. 6.3.4. We additionally discuss the trade-offs in 6.4.2, where it is demonstrated that while HybridSVD takes more time to train than PureSVD, it allows to achieve the same quality of recommendations at much lower rank values.

Reviewer Comment 13

Movielens does not instruct the users on what the stars mean. In user studies it is usually very clear what each rating means, but not in Movielens, so each user interprets this scale at his/her own will … I would recommend to discuss what happens with the proposed model if the threshold between positive and negative ratings is shifted.

Answer

Thank you for bringing our attention to this problem. Indeed, it is an important question. While our model allows to model ratings as ordinal values, the problem of a perceived scale and interpretation (especially on a personal level) is yet to be addressed. It is not something we have considered in this work. Within our model we stick to a simplified assumption that the rating scale is fixed and is the same for everyone. However, shifting the positivity threshold can be easily tested and we have actually performed such experiments. Our experiments show that lowering the threshold does improve benchmark in all tested models; however, the general conclusions on the performance of our model remain the same: our model is better at distinguishing between positively and negatively rated items, whichever threshold is chosen. We have added this observation in Sec. 5.4.3.

Reviewer Comment 14

I would recommend to include a comparison between the main existing frameworks and the presented one. I believe a comparison table would suffice.

Answer

Thank you for noting this. The comparison table is added in chapter 8.

Reviewer Comment 15

I would recommend to proofread the thesis. There are some language issues that should be addressed. A couple of examples: …

Answer

Thank you for helping to improve the text. We have proofread it and fixed the issues.

Reviewer: István Pilászy

Reviewer Comment 1

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]".

Answer

Thank you for noting this. The links referenced there actually contain the information about those authors. However, we agree, that this is confusing. The text was rephrased to make it explicit that we are not citing these authors or projects, but rather provide references for the works, where those authors and projects are mentioned together.

Reviewer Comment 2

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.

Answer

Thank you for bringing our attention to it. The entire subsection in the end of section 3.2 has been removed. The connection to the later sections is made more explicit.

Reviewer Comment 3

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.

Answer

We absolutely agree that tuning rank values (number of latent factors) is an important part of the experiment. However, the key idea of our experiments was to demonstrate the effect of positivity bias and show that good performance in terms of recommending relevant items does not necessarily lead to good performance in terms of avoiding irrelevant recommendations.

Moreover, different models have different optimal rank values and exhibit different behavior, when rank value is gradually changed in some range. Some models may quickly achieve their highest score at a lower rank and fail to improve further, other models may require much higher rank, however also provide much better evaluation scores (this effect can be observed on Figure 6.3, top-right graph, between FM and HybridSVD models). The difference becomes even more critical when comparing matrix- and tensor-based models, where latent spaces have very different structure. For example, increasing the rank value from 10 to 100 in SVD is not the same as increasing multilinear ranks from (13, 10, 2) to (130, 100, 2). Therefore, assigning the same randomly picked rank value to different models may not expose a clear picture of models' actual performance in comparison to each other. That's why we have computed models for rank 100 as you requested, and are ready to discuss the result; however, we have not included it into the thesis. On the other hand, we totally agree that having a picture of model's behavior with respect to a range of rank values would be an informative and useful addition to our work. Also note that in chapter 7, we do perform a grid search for estimating optimal hyper-parameters, including rank values. For every model the curves are obtained with the best found value of rank (multilinear rank), which in our opinion provides a fair comparison of models, including the models used in chapter 5.

Reviewer Comment 4

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

Answer

Every bar for PureSVD on the figure represents the value, corresponding to the best value of rank, i.e., at every sparsity level we have tuned the model in a range of rank values and picked the best one. We have now added this explanation to the caption of the figure.

Reviewer Comment 5

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).

Answer

Thank you for pointing us to this issue. Section 6.2.1 has been rewritten in order to make explanations more streamlined. We explicitly describe each required step for computing the model. We did not add pseudocode, because essentially it would be just a single line: computing singular triplets of an auxiliary matrix. In that sense the model is no more complicated than PureSVD. The only difference that Cholesky factors have to be precalculated. We have put efforts into making the text more understandable.

Reviewer Comment 6

In Section 6.2.2, please provide an example of creating such a similarity matrix …

Answer

We improved section 6.2, it now contains description of how similarity matrix can be computed with simple example of the features that can be used for that. We also provide an actual code for obtaining the result from Table 6.1 (the link is added to the table note).

Reviewer Comment 7

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?

Answer

Thank you for this question. Using a different similarity measure indeed affects the result. Generally it depends on the type of features. For example, as we state in Sec. 6.3.2, using Weighted Jaccard Index instead of cosine similarity for *ordered* collections of features helps improving the quality. However, the improvement in our case was not significant, typically within 1-2%.

Side information in our datasets indeed follows some long-tailed distribution. We, however, did not perform any additional experiments to find out, which distribution type (zipf or something else) it actually is. We also did not try to take into account the "long-tailedness" of side information, which can probably explain, why we were not observing a significant improvement, when switched

similarity measures. This can be an interesting direction for further research. Thanks for pointing this out.

Reviewer Comment 8

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?

Answer

Thank you for this question. Unfortunately, direct use of feature matrices instead of Cholesky decomposition is impossible as it wouldn't allow to solve the initial system of equations. As an example, solving this system requires having *square* matrices, otherwise the solution couldn't be derived due to inconsistencies in matrix operations. Moreover, we later need it to be invertable.

Nevertheless, there's one way to avoid Cholesky decomposition. If feature matrices are of a low rank, then it is possible to apply a trick, conceptually similar to Sherman-Woodberry-Morrison formula, and find matrix square root almost analytically (it only requires solving a system of that smaller size equal to the rank value). We have used that trick during the RecSys challenge in 2017, when there were more than million entities. This result, however, is not published yet and requires some additional research, thus we do not report it in the thesis.

Reviewer Comment 9

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

Answer

We agree that it would be an interesting experiment; however we haven't tried it yet.

Reviewer Comment 10

In eq (6.6) what is the matrix A ?

Answer

This was a typo. It is already corrected as we have improved the description of the HybridSVD method to make it more streamlined and transparent.

Reviewer Comment 11

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.

Answer

Thank you for bringing our attention to this issue. We have improved the description of the HybridSVD method to make it more streamlined and transparent.

Reviewer Comment 12

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

Answer

Thanks for noting this. We meant that it could be downloaded from their website. We have changed the wording accordingly.

Reviewer Comment 13

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?

Answer

We have added an explanation of how it can be computed based on a simple example with genres in the beginning of Sec 6.2. We also provide an actual code for obtaining the result from Table 6.1 (the link is added to the table note). This should help making the concept of similarity matrix clearer. The sparsity of the matrix in the Movielens case is 49% for ML1M and 51% for ML10M, so it is not really sparse. Thanks for bringing our attention to this. However, even in that case it was really easy to compute Cholesky decomposition, which takes around 30s for ML10M data on a modern laptop and is only required once unless new data arrives (in which case it is possible to

update the decomposition incrementally). Nevertheless, making similarity matrices sparser would definitely improve the computational performance. For example, excluding “Drama” from features pushes sparsity below 30% in both datasets and cuts the computation time down to approximately 20s. However, it does not improve the score. On the other hand, removing some non-informative features would probably help to achieve better performance. It requires a more rigorous research.

Reviewer Comment 14

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.

Answer

It was made intentionally. The top row demonstrates tuning of models; the bottom row compares models with the best configuration to each other. From the top row of this figure it can be seen that for Movielens alpha = 0.1 is the best choice for HybridSVD, while for BookCrossing it is alpha = 0.5. Therefore, in the bottom row we present the results for alpha0=0.1 on the left and for alpha=0.5 on the right.

Reviewer Comment 15

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?

Answer

Thanks for this question. Matrix L_S is triangular (by definition of the Cholesky decomposition), its inverse is also triangular. Its sparsity depends on the sparsity of the initial matrix; however it cannot have more than n(n+1)/2 non-zero elements (n is the size of initial matrix). What is more important, there’s no need to explicitly calculate the inverse. Triangular systems of equations can be efficiently solved with forward or backward substitution. We have added that note into the text of Sec. 6.2.3.

Reviewer Comment 16

Regarding the datasets / algorithms used in Section 6: although the author put a huge effort in finetuning Factorization Machines, I would like to see some experiments with other algorithms, or some experiments with other dataset. 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.

Answer

Thank you for noting this. We agree that having more datasets and more algorithms in comparison would help to present our ideas from a broader perspective.

Reviewer Comment 18

I would like to see some experiments with other algorithms … 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.

Answer

Thank you for noting these techniques. We absolutely agree that more rigorous comparison would provide a broader view on our method's performance. We’d like to add a few comments to this. The method by Gantner et al can be applied to both PureSVD and HybridSVD, as well as any other MF approach. In fact, it makes sense to do so, as the attributes are expected to be encoded by latent factors. This could potentially help to better handle the cold start scenario and would also provide an additional tool for analysing the differences of the learnt latent spaces from different models.

The concatenation approach, listed at the last position, potentially leads to an explosion of the latent feature space if the number of real features is high. In our model features are always aggregated

and do not require a separate latent space. Moreover, including both user attributes and item features becomes non-trivial with this approach.
Finally, we have selected FM algorithm because it has gained a certain popularity, has several very efficient and convenient to use implementations and does not require any additional modifications, allowing us to focus on its tuning. Moreover it generalizes a wide class of non SVD-based MF methods.

Reviewer Comment 19

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.

Answer

Thank you for this suggestion. In this section we tried to remove potential source of biases by only leaving highly-rated items and making their ratings binary. This leaves no confidence information for WRMF and would make its weighting function uniform. Nevertheless, we acknowledge the need for more rigorous comparison and we have added WRMF model into comparison in chapter 7, where we allow models to use all rating information. We have updated Figure 7.1 accordingly.

Reviewer Comment 20

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

Answer

We have placed the pseudocode of HOOI into Section 3.1.4, devoted to optimization algorithms for tensors.

Reviewer Comment 21

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.

Answer

Thank you for bringing our attention to this issue. We have changed the format of top-n evaluation graphs to vertical bar charts, which avoids overlapping. We find it easier to read than a long table with numbers.

Reviewer Comment 22

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?

Answer

Thank you for this suggestion. We have added the results of such experiment as a 4^{th} row on the figure. The HybridCoFFee model turns out to have the most vivid advantage over other models in this case. We have updated the text in Sec. 7.5 accordingly.

Reviewer Comment 23

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.

Answer

Thank you for noting this. Both issues are fixed now.

Reviewer Comment 24

I observed only a very few typos throughout the document…

Answer

Thanks for noting. We have fixed the issues you listed and additionally proofread the text.

Additional questions:

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

That's an interesting idea. Indeed, it is possible. It can be simply achieved by setting to zero those off-diagonal elements of the similarity matrix that correspond to popular items.

2. *Figure 7.1: for the BX dataset, CoFFee does not perform well for nDCG and nDCL. What could be the reason?*
   As we note in Sec 1.2.2, collaborative information is not always sufficient for learning representative patters even in the matrix case. Adding more dimensions in the way we do for the CoFFee model inevitably leads to even more extreme sparsity levels, which further worsens the problem. We can see from the figure, that the ratio or true positive to false positive is also not impressive in that case, which also supports our explanation. This is where the HybridCoFFee model shines, as it is designed to combat such problems with the help of side information and it clearly does. We have added one additional row to the figure, which corresponds to 3% fraction of ML10M dataset.

3. *For HybridSVD and PureSVD, do you handle user or item bias?*
   No, in our experiments biases are ignored. One of the ways to add such functionality is described in section 2.2.2.

4. *For HybridCoFFee: would it be possible to extend this method for 4 dimensional tensors?*
   Yes, absolutely. There is no theoretical limitation for doing this. However, going beyond 4 dimensions becomes problematic from the practical point of view due to the curse of dimensionality. More efficient formats like Hierarchical Tucker or Tensor Train should be employed in that case. This is, however, an unexplored territory and is a good direction for further research. We have added this remark into the conclusion of the thesis.

5. *What are the possible research directions to enhance the proposed algorithms further?*
   Two major directions are:
   1. Better warm start and cold start handling. Currently it directly transfers the result of folding-in from PureSVD. However, the presence of similarity matrices makes the folding-in problem more intricate. Deriving a more general hybrid variation of it seems to be both interesting and very practical.
   2. Excluding the need to hand-craft similarity matrix and make it a part of the optimization task, which does not require any input from the user. This may require some graph-based techniques.
   3. As noted above, adding more dimensions makes using Tucker decomposition infeasible and requires more efficient format. The best candidates are Hierarchical Tucker and Tensor Train decomposition.

6. *What are the most important open questions of recommender systems?*
   A very good overview of the most important challenges is nicely outlined in "Recommender Systems: Beyond Matrix Completion" by Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker; Communications of the ACM, November 2016, Vol. 59 No. 11, Pages 94-102.

   More philosophical matters are touched in "Algorithms Aside: Recommendation As The Lens Of Life" by Motajcsek, Tamas, et al; In Proceedings of the 10th ACM Conference on Recommender Systems, pp. 215-219. ACM, 2016. The quote that to some extent contains the whole concept is "*Life is the best recommender and the person is the Query*". It underlines the complicated nature of relations between us as humans, the real world around us and the digital world that affects our lives the way never before. It exposes a great challenge and reminds about a great responsibility.

Coming down closer to earth, what I would personally vote for is:

1. Debiasing.
   Data is missing not at random, we always have only a partial observation of a decision-making process and the mechanisms behind it are never exposed to us in full. This affects not only how and what algorithms learn, but also impacts the evaluation, leading to various biases at all levels. There are certain techniques already developed to date, however, some of them are far from being general and other raise some practical concerns of their applicability in real environments. Working towards a solution that builds a convincing representation of natural decision-making processes, allows to mimic or infer certain actions in a system and estimate their probability seems to be an important direction for active research.
2. Connecting offline and online evaluation results.
   There is still a dichotomy between academia and industry, business goals and quality metrics. Scientist rarely have access to real systems with active users, which prevents them from fully estimating capabilities of the newly proposed methods. Offline metrics not only have little connection to online performance, but sometimes can be even misleading. Probably, this also has something common with the previous point. Finding a proper way to conduct a fair comparison of algorithms even without testing them against real users would probably give a significant boost to the field.

7. *What is the future of matrix and tensor factorization methods in the field of recommender systems?*
   Generally, factorization methods provide one of the most efficient ways for solving the problem of generating good recommendations. Their flexibility and versatility finds application in various domains and allows to address many practical challenges. Some of the methods like iALS has become the default choice for newcomers, when setting up a new recommender system. Modern machine learning cloud platforms allow to build ready-to-use solutions with minimal efforts and even without infrastructure. Some factorization methods has already become a part of such offerings and their number will probably continue to grow as well as the number of the platforms.

   Taking into account the amount of work, devoted to the development and further improvement of these methods, it seems natural to assume that this work will continue and lead to new generalizations and discoveries, making solutions more feature-rich, easy to use, computationally efficient, etc. The world is changing rapidly, so it's hard to envision what other challenges are there waiting to be discovered and solved; however, some solutions are very likely to be based on factorization methods.

   Recommender systems started as a very practical field with a great focus on engineering part. On the other hand, an intersection with other fields of mathematics and connection to theoretical foundations has proven to lead to a fruitful symbiosis. I would expect a further coalescence of both practical tasks and theoretical investigations, bringing to existence some new methods with seemingly abstract ideas having a direct impact on real world applications.

Reviewer: Michael Thess

Reviewer Comment 1

[in chapter 6] the information about the construction of the side similarity matrices is a bit scarce. Given that the author acknowledges that this is a difficult topic I would expect to see more descriptions here. Especially, an example would be helpful.

Answer:

Thanks for pointing this out. Some parts of chapter 6 were rewritten to make reading more understandable. Additional examples of constructing similarity matrices were added. We also provide an actual code for obtaining the result from Table 6.1 (the link is added to the table note).

Reviewer Comment 2

I miss a continuative discussion of the work. Especially, how the proposed methods can be utilized for other applications than ratings (i.e. for web shops or search engines). Further, I miss a discussion about possible improvements and extensions.

Answer

Thank you for this remark. We have extended the conclusion part accordingly, emphasizing the applicability for different problems from various domains and also added discussion of potential issues in higher dimensions as well as possible ways to address them in future research.

Reviewer: Andrzej Cichocki

Reviewer Comment 1

A comparison with more recent state-of-the-art approaches would help to deliver a more complete picture on the applicability of the proposed methods. Extending experiments to several more datasets would also be beneficial and helpful in understanding of the method's performance

Answer

We agree that having more methods to compare with and more benchmark datasets to test on would help create a broader understanding of our methods' performance.

Reviewer Comment 2

Hybrid approach concerns with sparse similarity matrices only. An analysis of the possibility to operate on dense similarity matrices at large scale would be a good complementary part to the research.

Answer

Thank you for bringing this topic to the discussion. One possible way to deal with dense data is to use fast symmetric factorization, which in some sense is similar to Sherman-Woodberry-Morrison. It allows to find square root of a matrix of a special form almost analytically. We have tested this approach on large scale data with a million of entities and it demonstrated promising results. However, it is still a work in progress.

Reviewer Comment 3

Among 5 publications of the author, two are not yet published anywhere and exist only as preprints on except Arxiv

Answer

These papers were submitted to top conferences, including KDD, RecSys and NIPS during the year. Unfortunately they didn't pass the selection and will be submitted somewhere else.

Reviewer Comment 4

Presented methods are mostly applicable for the problems of order not higher than 3 or 4. It is still an open problem how to generalize them to higher orders as it would render both SVD and Tucker decomposition inapplicable.

Answer

Thank you for noting this. We mention this problem in the text and discuss potential ways of overcoming it in the conclusion.

Reviewer Comment 5

Another direction for further research is unification of data preprocessing which currently requires to construct similarity matrices and depends on the way it is done. More automated approach would help to move towards a more general solution.

Answer

We absolutely agree with this. This remains the topic for further research.

Reviewer Comment 6

Even though it was initially out of scope of the work and would require additional facilities, evaluation and comparison of the proposed methods on real users with the help of A/B testing would strengthen the contribution.

Answer

Thank you for noting this. Indeed, this is a very important problem, which we're planning to address in the future.

Reviewer: Alexander Tuzhilin

Reviewer Comment 1

One question that the candidate should think about and address at the thesis defense is this. Section 5.1.2 (and several other sections) state that the "Relevance Score denotes the likeliness of observing a certain (user, movie, rating) triplet." Is there any possible probabilistic interpretation of the "likeliness" notion for the Relevance Score? The candidate should comment on this probabilistic connection/interpretation and, possibly, explore it further.

Answer

Thank you for noting this. Looking forward to our discussion.

Reviewer: Dmitry Ignatov

Reviewer Comment 1

One recommendation for the defense. Please, if time allows, discuss relationship of your methods with Factorization Machines of SVD-like approaches with biases along with their added value.

Answer

Thank you for the suggestion. This is indeed an important and interesting topic. Looking forward to having the discussion on it.