Domain-Aware Grade Prediction and Top-n Course Recommendation

1. In terms of the problem formulation, is it right to not consider student interest and course content features? Is the model for top-n course recommendation biased towards 'popular' courses?
2. Some courses impose prerequisites and some courses differ from term to term (in terms of course structure, instructor, etc.). Can these factors be captured by some form of grouping or is there a need for explicit filters for these?
3. For testing, the authors use data from just one term while the training set contains data from the last 13 years. Do you think that the data would contain enough variation to make the results compelling?
4. The methodology suffers from cold-start issues as evidenced by the removal of new students from the test set. What are possible solutions to this?

Shreyas Udupa Balekudru

1. Not sure how course pre-requisite dependency is handled in this method?
2. Sometimes a course is taught by different professors in different semesters, and the course may be tailored differently based on the professor. E.g. Theoretical vs More hands-on. In that case, wouldn't taking into account the professor/lecturer also be important while creating recommendations?
3. Is metric like Recall@n sufficient? Course recommendation can shape a student's career, wouldn't taking into account factors like student's inclination, demand in industry, research scope, other factors related to the course subject be an important factor in recommending a course?

1. They do not take into account the different type of degree req. It is not the same for different institutions. Do you think it is robust enough?
2. Do you think a BPR model could be used here for ranking top-N courses?

-Dhruv SHarma

A Novel Recommender System for Helping Marathoners to Achieve a New Personal-Best

1. Types of runners: There are varied types of runners, depending on age, gender, or other user features. But the system only uses one non-personal best for all the users, which might be biasing the system towards the biggest category of runners. So, do you think that the clustering the users based on various attributes would have helped in building a better recommendation system.

2. Evaluation criterion: I have a concern regarding the evaluation criterion used in the paper. How do we judge if the system predictions were good? Is that when the system predicts the PB. of the user better than what the user performed?

Kriti Aggarwal

1. Picking the personal best is an important step in the recommender system since they train on <nPB, PB> pairs. In this paper, the pick the PB as the race with the fastest finish time, but there could be other races in which user finished maybe a little later but other factors like the average pace they mentioned is most uniform (with least deviation)?

2. Wouldn't the PB/nPB times also depend on the type of track? Do they take that into account? Was this the best way to choose PB vs nPB's?

3. They predict the new PB's and draw patterns in pace variations/ difference of PB's n PB's but there it seems there is no specific measure that tells if the predicted PB is accurate? Can we evaluate the PB correctly maybe by temporally dividing the dataset and see changes for some users who may have improved over time.? They don't really have test runners that can take their recommendations into account.

- Akanksha Grover

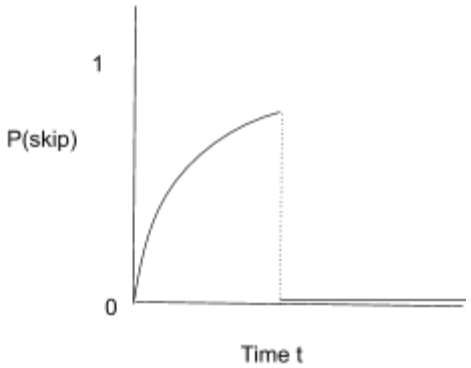
- I didn't see any suggested training plans described in the paper, despite that being what the authors said would be what best helps runners achieve a new PB. Were you able to find training plans from this paper, perhaps from the authors' webpage?
- Since the "results" of their paper can't actually be compared to real race results (which would still be difficult since they would have to control for the same race, race course, temperature, athlete's sleep/rest level, etc. in order to measure the new training plan), are there any other ways to evaluate the potential efficacy of the authors' methods?
- Stephanie Chen

Groove Radio: A Bayesian Hierarchical Model for Personalized Playlist Generation

1 - The process for deciding the candidate songs seem to be inefficient given that it requires retrieving artist which are similar to the current artist (given large number of artists). Authors did not go into much detail about this. Could this process be made more efficient?
2 - The authors did not talk about the runtime of their approach and given that the gradient calculation requires matrix inversion, I suspect that the runtime could be worse and method could be not that scalable. What are your thoughts?
3 - While defining the context features, authors mention that usage features' similarity is calculated using sigmoid function and meta-data features' similarity is calculated using cosine similarity. Do you have any insights into why different ways are used or is it just that whichever works best?

- By Rishabh Misra

1. How does the variational inference model allows prediction for under-represented data by borrowing information from sibling nodes with the same parent in the hierarchy? *(Kiran Kannar*)
2. The authors state that the model can be weighted on various similarity features - audio, meta-data, usage, popularity, etc. Could you clarify how? (Kiran Kannar)
3. Do you think the hierarchical classification is a complicated way of approaching the cold start problem for new artists/users? *(Kiran Kannar)*
4. It is notable that diversity is considered for practical considerations, but not in the results mentioned in the paper. Do you think AUC is a sufficient metric for comparison in this context? (*Kiran Kannar*)
5. Consider 30Music dataset. While it does not have skip data, (as lastfm does not scrobble user listening data if the user skips a song), it does have playtime of the song (over the actual length of the song). One way to improve this model is to simultaneously learn "skip" behaviour of uses, which could be as simple as a decay scheme where the probability to skip a song in increases till a threshold point, after which it can be assumed the user does not skip. *(Kiran Kannar)*



1. Is the sigmoid activation function the best way to model the probabilities / likelihood ?
2. Could there be significant overlap across the usage features got through bayesian model and the meta-data features. Would this result in any kind of double counting [but the AUC scores drop though]?
3. What could we do to avoid the rich get richer phenomenon shown in Figure 3 - Power law. There is a random exploration but is that sufficient?
4. Just like artists, could the users just have tags / favorite artists selected by themselves initially (when signing up to the service) to improve the AUC. This kind of meta-data of tag informations appears to be a very naive solution for both artists and users.
5. Apart from the probability what stops the model in just playing a repeated sequence of say 20 songs?

- Balasubramaniam Srinivasan

1. The authors focus on playlist generation, but why they care more about the artist not songs? Sometimes some artist has few songs which are popular but others are not popular, does it mean this model will not recommend this good song whose artists are not popular?
2. If some artist have different styles A, B, and the users previously like style A, style B most, do you think this system will recommend this artist to the user? Because the style vector must be different from all artists who focus on style A, and the artist who all focus on style B.
3. Sometimes people like to listen to quiet music at night and energetic music at morning, how to add this feature for this model?

   (Kuang Hsuan Lee)

By Rahul Dubey:

1. What's the reasoning behind using gamma distribution for hyper-priors, the tau parameters? AFAIK, gamma distributions are used in queueing models, or where time interval between successive events are relevant and to be modeled.

2. In the candidate selection step, it selects similar artists, does this provide enough diversity? Also, some users prefer to have diverse playlists, some don't. Is that taken into account here in user-specific parameters
3. In Groove dataset, a data point is considered a negative example, if the 2nd song is not played till the end, and skipped before its actual termination. Isn't that too strict a way to count negative examples? Many a times people skip the last few seconds of the song, but that may not imply that they don't want to listen to this song at all. Perhaps a threshold could be decided or probability could be assigned, rather than 0/1 label.

Gaze Prediction for Recommender Systems

1. The authors says that they don't model no fixation: they don't differentiate between different actions. Do you agree with the hypothesis? If they were to include this, what could be a good way to do that?
2. How would the model perform in cases where pages don't show constant recommendations or scrolling for example?
3. What other action-based or cursor-position based features be helpful in predicting the fixation in linear models?

- Tushar Bansal

1. The problem is modelled using a HMM in one of the algorithms. ALthough Si is calculated as the sum of posterior probabilities at each instance of time. In probability we add if the events are independent, but in HMM Xt is not independent of Xt-1. Do you think it could be modelled in anyother way?
2. Number of gaze users is 17. Do you think it represents the data correctly in such small number of users?

-Dhruv Sharma

Wednesday

Exploiting Food Choice Biases for Healthier Recipe Recommendation

1. This study seems to have been conducted more as a survey than an actual recommendation system. How do you think they could extend it to have

personalized recommendations (based on user preferences like vegetarian, vegans, etc.)?
2. As the study says, many users picked their recipe based on purely the images. However, images can be deceptive, as is often the case. How might they have better features to account for these cases?
3. The study has been conducted on ~100 undergraduate students in a particular age group, which might make the dataset very biased, in terms of dietary preferences and compulsions. Could you suggest a way to improve upon this without having personalizations?
4. Would it be beneficial to incorporate a learning model that generates the next recommendations based on user feedback (incorrect selection of less fatty item)? Or report which features made the user select that particular recipe, compared to its competitor?

Aditi Ashutosh Mavalankar

1. The paper does not consider user-specific features while making recommendations How would we extend this to include what types of food the user likes (eg. vegan) or what kind of healthy food they cook -low fats, low sugars etc.? If we use user recipe history as available on 'Allrecipes.com' under 'I made it' or user's rating of certain other recipes, maybe we could improve predictions and nudge users more towards choosing healthier options?
2. In some of the cases the results they show are better by just using image improvement features rather than their top 10 classification features. Made we wonder if users just select recipes by looking at images and we could make improvements using some specific image features?
3. Maybe we could also recommend users alternative portions or alternative ingredients for the recipes? This way they could make the recipes they like but with healthier ingredients.

- Akanksha Grover

1. The factors considered in determining the similarity of healthy/unhealthy recipes are unclear. Important aspects of recipe choices, other than already mentioned by others, such as amount of time required, budget, the complexity of the recipe do not seem to be considered.
2. By recommending healthier recipes based on their FSA score is assuming that the recipes are followed to the T. It is quite possible that the users modify the recipe to make it a tad bit unhealthier to very unhealthy. It is hard to track if the recommendation/nudge actually worked.

3. In the final study users reported that the healthiness of the recipes is important to them. Do you think along with inherent biases and image perception, it would be interesting to see how choices are made when it is explicitly stated that the recipe is healthier but still is very similar to what the user is looking for?
- *Sejal Shah*

Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors' Demonstration

1. The authors have treated the recommendations for articles published on the same day as independent of each other, but they do mention there may be correlations. How would one tweak the current model to incorporate that into the recommendation system?
2. It often happens that there are news items which are so popular that they are featured on the front page for days/weeks. In this case, the writing style matters less than it does for other articles that are not as important. Do you think this model will be able to handle that over time? If not, what do you propose in order for it to handle such events?
3. In the results section, they say that they have observed the results for Oct 1-9, out of which Oct 1-7 are Chinese holidays, and they go ahead to compare the recommendations for these 7 days with the other 2 days. Is this observation statistically significant (7 holidays and just 2 working days)?
4. If we look at Table 1, we see that the number of entries is skewed - the number of articles to be reviewed in each of the last two days is twice the number of articles to be reviewed in any of the other days (holidays). Is it not possible that the difference in recommendation is attributed to this change in the size of data?
    - Aditi Ashutosh Mavalankar

1. The author used a character-based CNN model to extract text features, since many languages does not have explicit "word" specification. Why not using different embedding models for different languages? Would that produce better results?
2. In the attention net, how does this method ensure that lambda_t learns the timeliness and lambda_m learns the model speciality?

3. This method discards the authors with low frequency (less than 3 times). But in realistic scenarios, articles of new authors may still be recommended by editors. Do you think all authors should be included during training?

- Siyu Jiang

1. While measuring the performance against the mentioned baselines in the A/B testing phase, it would be beneficial to also see how the DADM has performed on a test set from the manually created selections by professional editors that is used for training.
2. Do you think it would've been a more effective A/B test if the selections from the professional editors were also included as a comparative measure?
3. Would padding the shorter articles and clipping the longer articles to a length of 100 affect how the model learns/interprets the article?
4. There is no cold start problem where there is a huge set of demonstrations available, however for a new category/subcategory that comes with new buzz words, would the model function as effectively?
5. How often do you think the demonstration pool needs to be updated in the future? Can there be a quicker check than drop in article hits or losing app users/viewers?

- *Sejal Shah*

1. Character embedding doesn't consider text sequence. Would it be helpful to include sequential effect (e.g. introduction, conclusion) into the model?
2. The attention model is based on 2 assumptions, and it largely increases the complexity of the model. Would this lead to overfitting problem?
3. This model considers date as an important factor, but the authors only take experiment on 9 days (including holiday). I think it's better to have longer experiment.

- Zeng Fan

1. Does keeping a time profile imply that the model would keep recommending articles for some categories for longer time? Isn't this counter-intuitive as the model should promote newer content in every category.

2. How can we add more diversity to the recommendations? Recommender might be biased towards a single topic which might not be the best idea.
3. Authors raise an interesting point that they would like to study how their recommendations can impact further actions. How do you think we can analyze and somehow leverage this?

-Tushar Bansal

1. By considering news articles independent of each other, it's possible that two popular yet semantically very similar articles get picked. Wouldn't that affect diversity of the candidate set?

- Rahul Dubey

Multi-Modality Disease Modeling Via Collective Deep Matrix Factorization

1. The authors propose to initialize their model with a valid factorization of the original data matrix. Can you explain their initialization more clearly?
2. Can you compare their proposed method to neural factorization machine? - The motivations seem similar.
3. Is there a way to regularize or weight the model so that different modalities are represented more/less?
4. Can the authors impose more aggressive low-rank assumptions on the data to make the problem convex?

- Chester holtz

1. Do you think more ablation experiments should be done to better understand the method that handles modalities with missing objects? For example, experiments with a certain combination of different modalities could tell the importance of each modality.
2. In this specific task, how does other dimensionality reduction method performs? Are there any related works?
3. The user used linear, square and sigmoid activation function. Do you think other activation functions would perform better?

- Siyu Jiang

1. Why does the author do not compare their baseline with simple matrix factorization? Is it already taken into account?
2. Would applying PCA on individual modals and then concatenating them help?

- Digvijay Karamchandani

Personalized Key Frame Recommendation

1 - Authors divide each movie into L frames irrespective of the movie length and did not describe how they chose these frames. First, this would result in missing some potentially good frames and second, long movies be at more disadvantage. To mitigate this, I think selecting the number of key frames dynamically based on length of the movie would be more beneficial. What do you think?
2 - Authors model the text likelihood using RNN for each user and key frame in their training data. This data also contains negative samples (those where user sentiment is negative and those where user has not commented). Based on this, isn't the equation 8 wrong, given that it also models the text likelihood of the comments that are not available? If it is, how could we handle the cases where comments are not available?
3 - Do you think the model is scalable given that there are lots of components are parameters that are to be learn? Also, how could cold start problem be addressed?

- By Rishabh Misra

1. How representative are the frames commented on by users (or time-synchronized comments as expressed in the paper) of the actual importance of the frames? Is there some way to tell if different frames would be useful in different settings, e.g. which frame is more likely to get a user to click on a video vs. comment on it?
2. Are some of the baselines discussed in the paper at an inherent disadvantage, because these models are not intended to model video features? Are there baseline algorithms implemented on similar datasets that would be a better fit for this task?
3. Currently a positive example occurs when a user comments on a particular frame of the video. Are there other ways to collect data about important frames in the video that users may not have commented on? For example, would the thumbnails of videos the user is interested in also be valid training data?

- Rajiv Pasricha

Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks (Zhao et al. 2017)

1. Second order interactions have proved to be important based on the experiments on Yelp and Amazon (50k), similarly, will using neural FMs that capture higher order interactions improve recommending performance?
2. Given that meta graph similarity is already captured, do you think using ratings to obtain weighted HINs will improve model performance?

3. Will more metagraphs create more noise rather than improving the predictions?
4. Is there a quick way to learn the metagraphs in a dataset? This would be helpful considering the model proposed automatically selects features from significant metagraphs and rejects the ones that are not as good.

- *Sejal Shah*