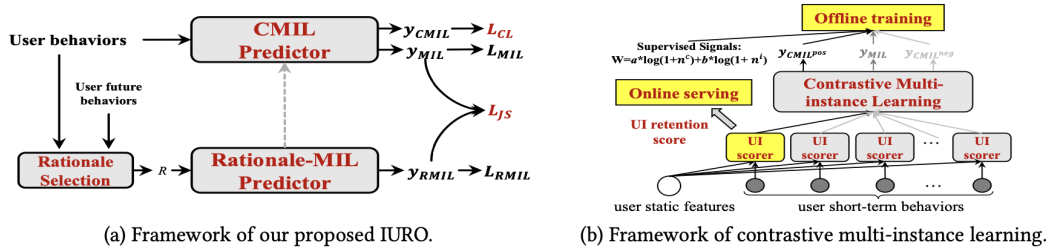
Анализ статьи. Бердышев Роман

1. Название статьи - Interpretable User Retention Modeling in Recommendation
   Авторы статьи - Rui Ding, Ruobing Xie, Xiaobo Hao, Xiaochun Yang, Kaikai Ge, Xu Zhang, Jie Zhou, Leyu Lin
   Год публикации - 2023
   Источник - 17th ACM Conference on Recommender Systems
2. Ссылка на статью - https://dl.acm.org/doi/10.1145/3604915.3608818. К сожалению, сейчас уже только платный доступ. Успел скачать pdf в самом начале. Искал еще в google scholar, но больше источников нет.
   Ссылка на гитхаб - https://github.com/dinry/IURO. К сожалению, гитхаб оказался пустой.
3. Краткое содержание
   Авторы пишут о том, что большинство рекомендательных моделей нацелены на увеличение метрик в моменте, например, CTR или время сессии, но такой подход не всегда хорош для бизнеса. Бизнес хочет, чтобы клиенты, как можно дольше пользовались их сервисом и метрикой показывающей это является retention - доля пользователей, вернувшихся через время. Проблема этой метрики в том, что она шумная и сложно оценить, что конкретно на нее влияет. Никто не подходил к настройке рекомендаций на retention и авторы предлагают собственный фрэймворк - interpretable user retention-oriented optimization (IURO), который находит и использует интерпретируемые факторы влияющие на пользовательский retention. Перед запуском модели авторы также провели количественное пользовательское исследование, в котором узнали, основные факторы влияющие на retention по мнению пользователей. Создатели модели сравнили в офлайне несколько бейзлайнов, а также протестировали лучшую модель на продакшене с помощью аб-теста и получили увеличение retention, при этом остальные метрики по типу CTR и время сессии не упали.
4. Основные тезисы
   - Почти никто не настраивал рекомендательные системы на долгосрочные

метрики, таргетируют на ctr и время сесси, но это может привести к кликбейту
- Если пытались настраивать рекомендации на долгосрочное воздействие, то использовали предсказывать последовательность сессий или уменьшать время между ними. Retention не использовали, потому что шумная метрика. Авторы придумали фрэймворк, который может объяснять, что побудило пользователей вернуться.
- Авторы исходят из предположения, что есть aha-moment, после которого пользователь с большой вероятностью вернется. Их мало, но они очень важные. Именно их и пытаются найти авторы
- Рекомендации настроенные на retention сложны в том, что у них есть два режима - офлайн и онлайн и они сильно отличаются между собой
- Факторы влияющие на retention выявить труднее, чем другое поведение - клики, просмотры и тд
- Данные для обучения более разреженные
- Фрэймворк состоит из двух частей CMIL - contrastive multi-instance learning и RMIL - rationale multi-instance learning
- Авторы провели количественное пользовательское исследование, где выяснили факторы влияющие на retention. Выделил положительные и негативные факторы
- Провели офлайн тесты 6 моделей и описали результаты
- Провели аб-тестирование в онлайне и получили увеличение метрик
- Проанализировали вклад дополнительных фичей в модель и перебрали гиперпараметры

5. Архитектура модели
   Модель состоит из двух частей - contrastive multi-instance learning и RMIL - rationale multi-instance learning. На рисунке 1(a) показана архитектура.



(a) Framework of our proposed IURO. (b) Framework of contrastive multi-instance learning.

Figure 1: Model overview of IURO. We attempt to explicit the intrinsic rationale of user retention via CMIL and RMIL.

В офлайне модель предсказывает вернется ли пользователь в течение нескольких дней, используя кумулятивные фичи(клики, просмотры) и сценарии взамодействия с айтемами.
CMIL модуль показан на рисунке 1(b) и он состоит из UI скореров, которые высчитывают скор для каждого короткого сценария(взаимодействия с айтемом), которые далее вместе с фичами юзера подаются в instance-level attention, который выдает свои скоры влияющие на retention. Contrastive заключается в том, что часть сценариев маскируется и на выходе аттеншена остаются только высоко-положительные и высоко-отрицательные сценарии(если нет contrastive, то это просто. MIL). Это хорошо влияет, так как авторы выделяют aha-moment - их мало и они очень сильно влияют на retention. Сам attention слой это трехслойная Feed Forward сеть. Лосс имеет такой вид

$$L_{CL} = \sum_{u \in U} max(0, |y_{MIL_u} - y_{CMIL_u^{pos}}| - |y_{MIL_u} - y_{CMIL_u^{neg}}| + m).$$

Чтобы использовать модель в онлайне используются кумулятивные фичи пользователя(просмотры, клики). Также ретеншн скор из UI скорера передается вместе с кумулятивными фичами для предсказания в онлайне и с помощью гиперпараметров регулируется соотношение между этими метриками.
Устройство RMIL аналогичное, только он обучается на будущих кликах пользователей из самый важных сценариев пользователя с предыдущего шага. С помощью RMIL стабилизируется выход модели при различных инициализациях. На финальном шаге с помощью дивергенции Йенсена - Шеннона авторы оптимизируют распределения из CMIL и RMIL

6. Результаты экспериментов

- Авторы провели пользовательский опрос, что им важнее всего в их рекомендациях в приложении(weChat) и почему они возвращаются.
    - **Положительные примеры**. Они ответили, что для большинства это персонализированные интересы, новости, различные шоу, но в зависимости от пользователя могут меняться. Большинство пользователей, сказали, что они вернуться к рекомендациям, если они будут позитивными, полезными и интересными, особенно важными

оказались расслабляющие и снимающие стресс рекомендации. Также важен дизайн страницы с рекомендациями, особенно для молодежи

- **Негативные примеры.** Самые плохие рекомендации - это неинтересные. Далее некачественные реки(насилие, фрод) или реклама. Далее недостаточное разнообразие и кликлбейты. Удивительно для авторов, что пользователи сталкивающиеся с повторными рекомендациями или не любит автора/контент имеет самый низкий отрицательный эффект.
  Они привели таблицу с ретеншеном и негативными причинами. И я не понимаю, как ее читать, есть причины, на которых ретеншн очень большой и даже больше 100%. Как это читать, авторы не объяснили

Table 1: Relative average user retention rates of different negative feedback reasons.

Explicit reasons of negative feedback	Next-day Retention	Next-three-day Retention
Not interested in items	35.38%	36.45%
Low-quality content (bloodiness, violence, fraud)	90.29%	91.78%
Advertising promotion	91.89%	93.06%
Poor diversity	97.33%	98.36%
Exaggerating titles	98.98%	99.05%
Don't like the author/content	102.29%	102.03%
Repeated recommendation	113.51%	108.07%

- Эксперименты.
  - Офлайн.
    Авторы использовали 6 моделей. Каждая последующая была усложнением предыдущей. Base MLP - просто подавали фичи пользователя и их действия за последние 3 дня в трехслойную MLP. IURO(AVG) - предполагали, что воздействие от всех айтемов одинаковое. IURO(MIL) - добавляли аттеншн и выделяли важные айтемы/сценарии. IURO(MIL+MSS) - выделяли кумулятивные фичи(количество кликов и просмотров) как фичи. IURO(CMIL+MSS) - добавляли contrastive learning.
    IURO(RCMIL+MSS) - добавляли механизм обоснования. Оценивали по метрике ROC-AUC и результаты в таблице:
    IURO(MIL) показала себя лучше Base MLP и IURO(AVG), что показывает, что разные сценарии по разному влияют на ретеншн пользователей и выделение важных сценариев помогает выделять aha-moment.
    IURO(MIL+MSS) показала себя лучше, то есть кумулятивные фичи могут

быть использованы для более точного предсказания. IURO(CMIL+MSS) показала себя еще лучше, показывая что contrastive learning лучше выделяет aha-moment. IURO(RCMIL+MSS) показала себя примерно также как IURO(CMIL+MSS), но этот способ лучше, так как он более устойчив к различным инициализациям.

Table 2: Results of offline user retention prediction (AUC).

Datasets	Base MLP	IURO (AVG)	IURO (MIL)	IURO (MIL+MSS)	IURO (CMIL+MSS)	IURO (RCMIL+MSS)
ZhihuRec	0.7180	0.5838	0.8269	0.8312	0.8437	**0.8445**
Industry	0.6827	0.6732	0.7209	0.7255	0.7291	**0.7301**

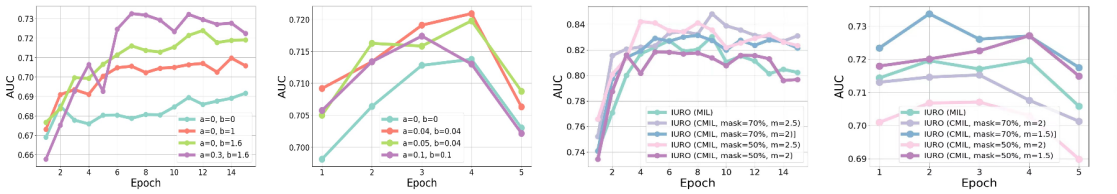
- Онлайн

  В онлайне авторы провели АБ-тест на 3 млн пользователях, где сравнивали контрольную группу с обычной моделью, нацеленной на ctr и тестовую группу, где совмещались ретеншн скоры UI скорера и ctr скоры по формуле. В эксперименте авторы сфокусировались на ретеншне на следующий день и на трехдневном ретеншнене, так как это основные метрики ретеншена для их сервиса. Результаты в таблице
  На основе этих данных авторы пишут, что изменения в тестовой группе есть и они подтверждают интерпретируемость ретеншена пользователей и работоспособность их подхода. Также авторы проверили, что другие метрики, такие как CTR и продолжительность сессии не упали в тестовой группе. При том они выделили отдельный сегмент hit items - это те рекомендации, которые выдала их модель и юзеру они понравились. И в тестовой группе ctr на этом сегменте оказался выше.

Table 3: Online A/B test on a widely-used industrial article recommendation feed.

Model	Next-day Retention	Next-three-day Retention
IURO(RCMLI+MSS)	+0.76%	+0.65%

Авторы дополнительно проанализировали использование гиперпараметров в модели.



(a) Impacts of SS on ZhihuRec. (b) Impacts of SS on Industry. (c) Results of CMIL on ZhihuRec. (d) Results of CMIL on Industry.

Figure 2: Exploration on interpretable user retention-oriented optimization.

Использование кумулятивных фичей улучшало предсказания модели. Это видно на графиках а, б. При занулении гиперпараметров a,b эти фичи не используются.
Авторы использовали разные стратегии в contrastive learning - делали разные маски, то есть брали x% топ аттеншн скоров и x% самых низких аттеншн скоров и обучали модель на них. И видно, что с использованием маски предсказания улучшаются.

Кроме этого авторы выделили сегменты рекомендаций в зависимости от ретеншн скоров.

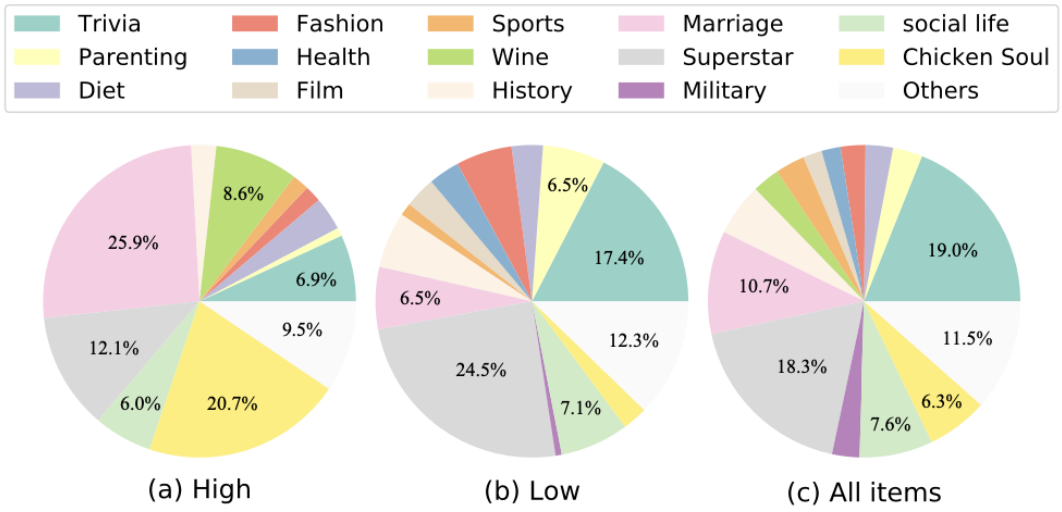


Figure 3: Category distribution of a user group's historical click behaviors with different UI retention scores.

И получили вывод, что распределение ctr по категориям среди низких аттеншн скоров и всех скоров очень похожее, а вот распределение по категориям среди больших ретеншн скоров не похоже на все айтемы. Это важный инсайт, который можно дальше использовать при развитии рекомендательных систем.

7. Анализ статьи
   Сильные стороны:
   - наличие перебора гиперпараметров
   - описали новизну решаемой проблемы, привели ссылки на источнике, где решалась похожая проблема
   - придумали новый фрэймворк
   - протестировали офлайн и онлайн
   - описание неудачных экспериментов
   - дополнительный анализ модели
   - сегментация рекомендаций и интересные инсайты
   Слабые стороны:
   - нет открытого кода для воспроизведения модели
   - не использованы открытые датасорсы, а использованные нигде не предоставлены
   - не было простых бейзлайнов

- в аб-тесте не указаны статзначимости. Неясно насколько можно верить результатам. Так retention шумная метрика, то диспресия у нее может быть большая
- недостаточное описание архитектуры модели
- отсутсвие дальнейших шагов
- нет бейзлайнов из разных семейств алгоритмов
- пишут про долгосрочный ретеншн, а таргетируются на одно-, трехдневный ретеншн

8. Мне кажется, вклад этой статьи может быть достаточно большой, так как авторы подсветили проблему, что нужно настраивать рекомендации на долгосрочную перспективу, а не только в моменте выдавать популярные рекомендации. До них задумывались о такой проблеме, но они подобрали правильную метрику и смогли выявить факторы, которые на нее влияют.
9. Авторы не предлагают никаких дальнейших шагов
10. Я бы предложил заполнить репозиторий на гитхаб, чтобы можно было воспроизвести результаты и лучше разобраться в архитектуре. Кроме этого, попробовать настроить модель на более долгосрочный ретеншн, не только на 3 дня вперед. Возможно разработать разные модели для новичков и старичков, так как проблема холодного старта тут тоже может быть, но про нее авторы ничего не писали