

## **МЕТРИКИ И МЕТОДОЛОГИЯ СРАВНИТЕЛЬНОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ ПЕРСОНАЛЬНЫХ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ**

Е.А. Злобин, А.С. Кравченко

ФГБОУ ВО «Воронежский государственный лесотехнический университет  
имени Г.Ф. Морозова»

Аннотация. В статье систематизированы метрики оценки эффективности алгоритмов персональных рекомендательных систем: метрики точности ранжирования, ошибки предсказания, покрытия, разнообразия и новизны. Проанализированы протоколы валидации (hold-out, кросс-валидация, leave-one-out) и влияние проблемы холодного старта на достоверность оценки. Обоснована необходимость многокритериального подхода при сравнительном анализе рекомендательных алгоритмов.

Ключевые слова: рекомендательные системы, метрики оценки, кросс-валидация, холодный старт, сравнительный анализ.

## **METRICS AND METHODOLOGY FOR COMPARATIVE EVALUATION OF THE EFFECTIVENESS OF PERSONALIZED RECOMMENDER SYSTEM ALGORITHMS**

E.A. Zlobin, A.S. Kravchenko

Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The article systematizes performance evaluation metrics for personalized recommender system algorithms, including ranking accuracy, prediction error, coverage, diversity, and novelty metrics. Validation protocols (hold-out, cross-validation, leave-one-out) are analyzed along with the impact of the cold start problem on evaluation reliability. The necessity of a multi-criteria approach in the comparative analysis of recommender algorithms is justified.

Keywords: recommender systems, evaluation metrics, cross-validation, cold start, comparative analysis.

Персональные рекомендательные системы повсеместно используются в современных сервисах, а выбор конкретного алгоритма невозможен без количественного сравнения их эффективности. При этом оценка рекомендаций принципиально отличается от классических задач машинного обучения: система формирует ранжированный список с учётом индивидуальных предпочтений пользователя. Это обуславливает необходимость специализированных метрик, которые традиционно делятся на три группы: точность ранжирования, ошибка предсказания явных рейтингов и свойства самого списка (покрытие, разнообразие) [1].

Наиболее распространённой постановкой задачи является формирование топ-N списка рекомендаций. Качество такого списка оценивается путём сравнения с множеством объектов, действительно релевантных для пользователя (тестовая выборка). Данная группа метрик ориентирована на бинарную релевантность: объект либо присутствует в тестовом множестве пользователя, либо отсутствует.

$Precision@k$  определяется как доля релевантных объектов среди первых  $k$  позиций сформированного списка:

$$Precision@k = \frac{|R_k \cap T|}{k},$$

где  $R_k$  - множество объектов в первых  $k$  позициях рекомендации,  $T$  - множество релевантных объектов пользователя в тестовой выборке. Метрика отвечает на вопрос: «Какова вероятность того, что случайно выбранный объект из топ- $k$  окажется полезным?»

$Recall@k$  отражает долю найденных релевантных объектов от общего их числа:

$$Recall@k = \frac{|R_k \cap T|}{|T|}.$$

$Recall$  характеризует полноту выдачи: при фиксированном размере тестового множества рост  $Recall$  свидетельствует о способности алгоритма находить большее число значимых объектов. Взаимосвязь  $Precision$  и  $Recall$  носит компромиссный характер: увеличение длины списка  $k$ , как правило, повышает  $Recall$ , но снижает  $Precision$ .

F1-мера объединяет  $Precision$  и  $Recall$  через гармоническое среднее:

$$F1@k = \frac{2 \cdot Precision@k \cdot Recall@k}{Precision@k + Recall@k}.$$

Когда важна не только релевантность объекта, но и его позиция в выдаче, используют метрику  $NDCG@k$  (Normalized Discounted Cumulative Gain):

$$DCG@k = \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}, NDCG@k = \frac{DCG@k}{IDCG@k},$$

где  $rel_i$  обозначает степень релевантности объекта на позиции  $i$ , а  $IDCG@k$  соответствует идеальному ранжированию. Логарифмический дисконт здесь играет ключевую роль: чем дальше объект от начала списка, тем меньше его вклад в итоговую оценку. Значения NDCG лежат в интервале  $[0; 1]$ . В отличие от Precision и Recall, эта метрика работает с градированной релевантностью, например с пятибалльными оценками, что позволяет применять её одновременно к данным с явной и неявной обратной связью [2].

Для усреднения результатов по всем пользователям тестовой выборки применяется MAP (Mean Average Precision):

$$AP = \frac{1}{|T|} \sum_{k=1}^n Precision@k \cdot rel_k, MAP = \frac{1}{|U|} \sum_{u \in U} AP_u,$$

где  $U$  - множество пользователей в тестовой выборке.

Ещё одна полезная метрика, MRR (Mean Reciprocal Rank), учитывает только позицию первого релевантного объекта:

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u},$$

где  $rank_u$  - позиция первого подходящего объекта в рекомендациях пользователя  $u$ . MRR хорошо показывает себя там, где пользователю достаточно одного результата, скажем при подборе единственного учебного материала или ответа на запрос.

На практике выбор метрики зависит от конкретной задачи. В длинных лентах рекомендаций, характерных для новостных агрегаторов или музыкальных сервисов, чаще ориентируются на Recall и NDCG. Для коротких подборок из 5-10 товаров более показательны Precision и MRR.

Если задача сводится к предсказанию явного рейтинга, например оценки фильма по шкале от 1 до 5, на первый план выходят регрессионные показатели. Два основных из них:

Средняя абсолютная ошибка (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |r_i - \hat{r}_i|,$$

Среднеквадратичная ошибка (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - \hat{r}_i)^2},$$

где  $r_i$  - фактический рейтинг,  $\hat{r}_i$  - предсказанный,  $n$  - число оценок в тестовой выборке.

Разница между ними принципиальна. За счёт квадратичного штрафования RMSE сильнее реагирует на крупные отклонения. Если для сервиса критично не допускать серьёзных промахов в прогнозе, предпочтительнее RMSE. Обе метрики активно использовались в ходе Netflix Prize и до сих пор остаются стандартом при бенчмаркинге рекомендательных моделей [3].

Иногда встречается нормализованный вариант, NMAE, в котором ошибка делится на диапазон шкалы:

$$NMAE = \frac{MAE}{r_{max} - r_{min}}.$$

Такая нормализация удобна, когда нужно сопоставить результаты, полученные на датасетах с разными шкалами. Наряду с этим применяется коэффициент детерминации  $R^2$ , отражающий долю дисперсии целевой переменной, объяснённую моделью. Значения  $R^2$ , близкие к единице, говорят о хорошем качестве предсказания. Отрицательные значения сигнализируют о том, что модель уступает банальному прогнозу средним значением.

Если оценивать рекомендации только по точности, система рискует скатиться к выдаче однотипных объектов из узкой части каталога. Пользователь, которому постоянно показывают одни и те же популярные фильмы или товары, быстро теряет интерес. Поэтому наряду с метриками точности используют дополнительные критерии, отражающие качество рекомендаций с точки зрения пользовательского опыта и экономики платформы [4].

Покрытие (Coverage) показывает, какую долю каталога система в принципе способна рекомендовать:

$$Coverage = \frac{|\{i : \exists u \ i \in R(u)\}|}{|I|},$$

где  $I$  - полный каталог объектов,  $R(u)$  - список рекомендаций для пользователя  $u$ . Низкое покрытие обычно указывает на популярностный уклон: алгоритм концентрируется на небольшом наборе часто оцениваемых объектов и практически не затрагивает длинный хвост каталога. Для коммерческих платформ это оборачивается упущенными продажами нишевых товаров.

Разнообразие (Diversity) оценивает, насколько различаются объекты внутри одного рекомендательного списка. Распространённый способ основан на попарном сходстве:

$$Diversity = 1 - \frac{2}{k(k-1)} \sum_{i < j} sim(item_i, item_j),$$

где  $sim$  - функция сходства между векторами признаков объектов. В качестве такой функции могут выступать косинусная близость, корреляция Пирсона и другие меры. Чем ниже среднее попарное сходство, тем выше разнообразие выдачи.

Высокое разнообразие означает, что объекты в списке существенно различаются по атрибутам.

Новизна (Novelty) отражает степень неожиданности рекомендации для конкретного пользователя. Формально новизна объекта  $i$  для пользователя  $u$  может быть определена через его популярность:

$$Novelty(i) = -\log_2 p(i),$$

где  $p(i)$  - вероятность взаимодействия с объектом  $i$  в обучающей выборке. Редкие, малоизвестные объекты обладают высокой новизной; рекомендации, дублирующие общеизвестные хиты, - низкой.

Неожиданность (Serendipity) является наиболее комплексным показателем: рекомендация считается серендипитной, если она одновременно релевантна и не может быть тривиально выведена из известного профиля пользователя. Формализация серендипитности остаётся предметом дискуссий; один из подходов определяет её как долю релевантных объектов, отсутствующих в расширенном профиле пользователя.

Достоверность сравнительного анализа определяется не только выбором метрик, но и схемой разделения данных на обучающую и тестовую части. Неадекватный протокол валидации способен привести к завышенным оценкам качества и неверным выводам о превосходстве одного алгоритма над другим [5].

Hold-out. Исходный набор взаимодействий каждого пользователя разделяется на обучающую (обычно 80 %) и тестовую (20 %) части. Метод прост в реализации и требует однократного обучения модели. Однако результаты существенно зависят от конкретного случайного разбиения: при малом числе взаимодействий пользователя случайное попадание или непопадание ключевого объекта в тестовую выборку способно исказить оценку. Для снижения дисперсии процедуру повторяют с несколькими случайными разбиениями и усредняют результаты.

k-кратная кросс-валидация. Данные разбиваются на k непересекающихся блоков (фолдов); модель обучается на k-1 блоках и тестируется на оставшемся. Процедура повторяется k раз таким образом, чтобы каждый блок побывал в роли тестового. Итоговые метрики усредняются по всем фолдам. Типичные значения k = 5 или 10. Кросс-валидация позволяет снизить дисперсию оценки и задействовать данные полнее, чем при простом разбиении. Обратная сторона состоит в том, что вычислительная стоимость возрастает в k раз, поскольку модель приходится обучать заново на каждом фолде. На датасетах с миллионами взаимодействий такие затраты часто оказываются неоправданными.

Протокол leave-one-out (LOO) устроен иначе. Из взаимодействий каждого пользователя в тестовую выборку отводится ровно одно, а всё остальное идёт в обучение. Такой подход распространён в задачах ранжирования, где нужно проверить, способен ли алгоритм предсказать единственное следующее действие пользователя. Обучение при этом получает максимально возможный объём данных. Платой за это становится высокая вычислительная нагрузка, особенно когда число пользователей велико.

Отдельного внимания заслуживает временное разделение (temporal split). В тех сценариях, где важна последовательность событий, данные разбиваются по временной метке. Взаимодействия, совершённые до момента  $t_0$ , составляют обучающую выборку, а всё, что произошло после  $t_0$ , отходит в тестовую. Такое разделение исключает утечку информации из будущего и ближе всего к тому, как система будет работать в реальных условиях.

Отдельную методологическую сложность представляет оценка качества рекомендаций в условиях холодного старта [6]. Выделяют три основных варианта данной проблемы: холодный старт нового пользователя (отсутствие или минимальное число взаимодействий), холодный старт нового объекта (отсутствие оценок со стороны сообщества) и холодный старт системы в целом (начальный этап функционирования сервиса).

Для практической реализации сравнительного анализа в рамках разрабатываемого программного обеспечения рассмотрены существующие библиотеки и фреймворки.

Surprise (Python) - легковесная библиотека, предоставляющая набор алгоритмов коллаборативной фильтрации (SVD, KNN, Slope One) и стандартные метрики (RMSE, MAE, Precision, Recall, F1). Поддерживает кросс-валидацию и подбор гиперпараметров через GridSearch. Ограничения: отсутствие метрик ранжирования (NDCG, MRR) и контент-базированных моделей.

RecBole - унифицированный фреймворк, включающий более 90 моделей (коллаборативные, контент-базированные, последовательные, графовые) и расши-

ренный набор метрик, в том числе NDCG, MRR, Hit Rate, Coverage. Обеспечивает стандартизированный конвейер «загрузка данных, обучение, оценка, сравнение».

LensKit (Java) - исследовательский фреймворк, ориентированный на воспроизводимость экспериментов. Поддерживает гибкие протоколы оценки, включая временное разделение, и предоставляет детализированные отчёты на уровне отдельных пользователей. Экосистема библиотеки менее активно развивается в последние годы.

Проведённый анализ показывает, что объективное сравнение алгоритмов персональных рекомендательных систем требует многокритериального подхода, включающего метрики точности ранжирования, ошибки предсказания, покрытия и разнообразия. Выбор протокола валидации должен учитывать объём данных, временную природу взаимодействий и специфику предметной области. Стратификация по степени холодного старта позволяет получить более дифференцированную картину эффективности методов и обосновать выбор гибридных стратегий.

### Список литературы

1. Кульшин, Р. С. Рекомендательные системы в цифровых сервисах / Р. С. Кульшин, А. А. Сидоров, П. В. Сенченко // Электронные средства и системы управления: материалы докладов Международной научно-практической конференции. – 2023. – № 1. – С. 352-354. – EDN KSOUYC.
2. Маковейчук, К. А. Системное моделирование предпочтений в рекомендательных системах с использованием интерпретируемых подходов машинного обучения / К. А. Маковейчук // Дистанционные образовательные технологии: сборник трудов X Международной юбилейной научно-практической конференции. – Симферополь. – 2025. – С. 329-332. – EDN EGIAGA.
3. Соколов, М. И. О разработке рекомендательной системы для электронной коммерции / М. И. Соколов, А. И. Долженко // Научный вектор: сборник научных трудов. Том Выпуск 9. Под научной редакцией Е.Н. Макаренко. – Ростов-на-Дону. – 2023. – С. 263-267. – EDN QCYLPZ.
4. Головинский, П. А. Нечеткая рекомендательная система с холодным стартом для выбора траектории обучения / П. А. Головинский, А. О. Шаталова // Проблемы управления. – № 6. – 2023. – С. 33-41. – EDN HVRWRH.
5. Шагеев, Д. А. Оценка уровня конкурентоспособности российских банков при помощи модифицированного метода анализа иерархий / Д. А. Шагеев, Д. А. Заболотнев // Моделирование систем и процессов. – 2025. – Т. 18, № 3. – С. 121-139. – DOI 10.12737/2219-0767-2025-18-3-121-139. – EDN НКТҮНҮ.

6. Сумин, В. И. Использование ситуационного моделирования в разработке систем принятия решений для сложных организационных систем / В. И. Сумин, А. С. Кравченко, А. В. Толкачев // Моделирование систем и процессов. – 2024. – Т. 17, № 3. – С. 71-79. – DOI 10.12737/2219-0767-2024-69-77. – EDN FBBHJO.

## References

1. Kulshin R.S. Recommender systems in digital services / R.S. Kulshin, A.A. Sidorov, P.V. Senchenko // Electronic means and control systems: proceedings of the International scientific-practical conference. – 2023. – No. 1. – P. 352-354. – EDN KSOUYC.
2. Makoveychuk K.A. System modeling of preferences in recommender systems using interpretable machine learning approaches / K.A. Makoveychuk // Distance educational technologies: proceedings of the X International jubilee scientific-practical conference. – Simferopol. – 2025. – P. 329-332. – EDN EGIAGA.
3. Sokolov M.I. On the development of a recommender system for e-commerce / M.I. Sokolov, A.I. Dolzhenko // Scientific vector: collection of scientific papers. Vol. 9. Under the scientific editorship of E.N. Makarenko. – Rostov-on-Don. – 2023. – P. 263-267. – EDN QCYLPZ.
4. Golovinsky P.A. Fuzzy recommender system with a cold start for selecting a learning trajectory / P.A. Golovinsky, A.O. Shatalova // Control problems. – No. 6. – 2023. – P. 33-41. – EDN HVRWRH.
5. Shageev D.A. Assessing the competitiveness level of Russian banks using a modified analytic hierarchy process / D.A. Shageev, D.A. Zabolotnev // Modeling of systems and processes. – 2025. – T. 18, No. 3. – P. 121-139. – DOI 10.12737/2219-0767-2025-18-3-121-139. – EDN HKTYHV.
6. Sumin V.I. Using situational modeling in the development of decision-making systems for complex organizational systems / V.I. Sumin, A.S. Kravchenko, A.V. Tolkahev // Modeling of systems and processes. – 2024. – Т. 17, No. 3. – P. 71-79. – DOI 10.12737/2219-0767-2024-69-77. – EDN FBBHJO.