УДК 001.891.5 DOI: http://doi.org/10.25728/pu.2024.4.4

ОБОБЩЕННАЯ МЕТРИКА ОЦЕНКИ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ ЭНТРОПИЙНОГО МЕТОДА[#]

Р. С. Кульшин*, А. А. Сидоров**

Томский государственный университет систем управления и радиоэлектроники

*✉ roman.s.kulshin@tusur.ru, **✉ anatolii.a.sidorov@tusur.ru

Аннотация. Рассматривается задача формирования интегрального показателя для оценки эффективности алгоритмов рекомендательных систем, который был создан путем объединения отдельных метрик с использованием энтропийного метода. Работа основывается на исследовании в качестве базы для тестирования набора из 12 алгоритмов, с одной стороны, и трех наборов данных, с другой, для каждой комбинации которых были рассчитаны отдельные критерии, используемые в практике оценки рекомендательных систем. Результаты исследования свидетельствуют о том, что полученный интегральный показатель является эффективным инструментом оценки работы алгоритмов рекомендательных систем. Показано, что качество работы алгоритмов различается в зависимости от размера и иных базовых характеристик набора данных. Обобщенная мера может быть использована для разработки более эффективных алгоритмов, их ансамблей, оптимизации гиперпараметров и улучшения качества рекомендаций.

Ключевые слова: рекомендательные системы, интегральный показатель, алгоритмы, метрики, наборы данных.

ВВЕДЕНИЕ

В рамках гибридного мира, когда реальные общественные отношения тесным образом переплетаются с цифровыми практиками коммуникаций и экономических активностей, рекомендательные системы становятся востребованными сервисами среди различных бизнес-акторов, как предлагающих на рынке различные виды продукции, так и потребляющих их. Так, в условиях постоянного роста объемов данных получатели различных благ сталкиваются с необходимостью эффективного поиска и фильтрации информации [1], а их продуценты ищут способы повышения эффективности маркетинговой деятельности, направленной на стимулирование потребления производимых товаров и услуг. В обозначенном контексте рекомендательные системы используются для предоставления персонализированных рекомендаций в различных бизнес-сегментах: от онлайн-покупок до потокового воспроизведения музыки и фильмов.

В сфере электронной коммерции рекомендации товаров позволяют покупателям находить продукты, наиболее соответствующие их интересам и предпочтениям. Рекомендательные системы анализируют данные о предыдущих покупках, просмотрах товаров [2] и поведенческих паттернах и на этой основе предлагают наиболее релевантные варианты. На стриминговых платформах для прослушивания музыки и просмотра фильмов рекомендации играют ключевую роль в обеспечении удовлетворения интересов пользователей: на основе анализа предпочтений относительно жанров, исполнителей, режиссеров, а также посредством учета рейтингов и отзывов предоставляются персонализированные списки воспроизведения и рекомендации контента [3].

Технологии, лежащие в основе рекомендательных систем, включают в себя машинное обучение, алгоритмы коллаборативной фильтрации и контент-фильтрации. Многие из них также используют нейросетевой подход [4] для более точного анализа и предсказания предпочтений пользователей. Востребованность рассматриваемых программных сервисов и продуктов порождает появ-

[#] Работа выполнена в рамках государственного задания Минобрнауки России; проект FEWM-2023-0013.

ление большого количества алгоритмов, метриками оценки эффективности которых являются два параметра: время и память, характеризующиеся асимптотическим ограничением. Самым распространенным методом оценки алгоритмов является Big O [5]. Нотация Big O используется для описания сложности алгоритмов и позволяет сравнивать их эффективность на основании данных одного типа.

В тоже время, в концепции рекомендательных алгоритмов существует множество других метрик, характеризующих результаты их работы. Перечень критериев оценки довольно обширен и включает в себя как параметры, характерные для методов машинного обучения в целом, так и специализированные метрики рекомендательных алгоритмов [6]. Они измеряют точность, полноту, разнообразие рекомендаций, надежность, понятность, удовлетворенность пользователей, покрытие каталога, скорость обработки запросов и другие характеристики. Каждая метрика имеет свои преимущества и ограничения в зависимости от контекста применения и целей системы рекомендаций.

Использование разнообразных метрик позволяет получить более полное представление о различных обобщенных аспектах рекомендательных алгоритмов, не имеющих, как правило, общеразделяемого научным и профессиональным сообществом смысла (производительности, результативности, эффективности, качестве) и пригодности алгоритмов для конкретных задач и аудиторий.

В обозначенной ситуации при сравнении алгоритмов и выборе того или иного из них для решения конкретной задачи возникают следующие трудности:

• Существует большое количество метрик оценки алгоритмов, положенных в основу рекомендательных систем, каждая из которых подходит для оценки разных аспектов рекомендаций. Это может привести к неоднозначным результатам, когда различные метрики указывают на разные алгоритмы как на лучшие.

• Некоторые метрики могут быть взаимосвязаны или зависеть друг от друга. Улучшение значения одной метрики может привести к ухудшению значения другой. Это усложняет принятие решения о том, какой из алгоритмов является более подходящим.

• Значения метрик могут варьироваться в зависимости от контекста. Например, точность может отличаться в различных сферах применения.

Создание единой метрики оценивания рекомендательного алгоритма является необходимым шагом для обеспечения объективности, сравнимости и улучшения качества рекомендаций. Это позволяет стандартизировать оценку эффективности алгоритмов, выбрать наиболее подходящий для конкретной задачи и оптимизировать использование ресурсов.

Основная гипотеза работы заключается в том, что все алгоритмы формирования рекомендаций ведут себя по-разному в зависимости от контекста данных, размера наборов данных и прочих характеристик среды формирования рекомендаций. Для того чтобы избежать трудозатратного эмпирического подбора алгоритма со сравнением множества метрик под конкретную ситуацию, предлагается создать суперкритерий оценки рекомендательных алгоритмов.

1. МЕТОДОЛОГИЯ

1.1. Интегральный показатель как обобщенная мера

В качестве единого критерия предлагается использовать интегральный показатель, представляющий сводную величину, объединяющую в себе несколько отдельных релевантных метрик для измерения сложной синтетической конструкции [7–10]. На сегодняшний день интегральные показатели используются в различных областях для оценки:

– производительности компаний, рынков и экономики в целом [11–13];

– состояния здоровья пациента [14–16];

– состояния окружающей среды [17–19];

– рисков в различных областях [20–22];

– качества образования и успеваемости студентов [23–25];

– и др.

Они имеют свои преимущества и недостатки. Так, к преимуществам интегральных показателей можно отнести то, что они:

• позволяют объединить несколько отдельных индикаторов или переменных в обобщенную метрику, что делает их удобными для анализа и сравнения [7, 9, 10, 24];

• позволяют проводить анализ на базе логического содержательного объединения переменных, связанных с определенной областью или темой [9, 10, 13];

• могут быть использованы во многих областях и для решения различных задач, что делает их универсальными инструментами для анализа и прогнозирования [9, 13].

Говоря о достоинствах интегральных показателей, нельзя не отметить и некоторые недостатки данной концепции:

• узким местом рассматриваемой методологии является вопрос присвоения весов отдельным индикаторам, что порождает возможность манипулирования в рамках обоснования того или иного решения, принятого на основе результатов расчетов обобщенной меры [7, 24];

• использование интегральных показателей может быть ограничено доступностью данных, требуемых с точки зрения содержательного обоснования структуры суперкритерия [7, 13].

Несмотря на указанные недостатки, интегральные показатели могут быть эффективными, так как возможность их применения во многих контекстах наделяет их свойством гибкого инструментария, способного адаптироваться под различные задачи и условия.

Само по себе применение методологического подхода, основанного на исчислении интегрального показателя, представляется достаточно дискуссионным, что породило появление в научном сообществе двух лагерей – его противников и сторонников. Несмотря на разнообразие точек зрения и спорных моментов (например, в части отбора первоначального перечня критериев, выбора методов определения весовых коэффициентов и агрегации, способов нормализации показателей, измеряемых в различных диапазонах и шкалах), обобщенная мера представляет достаточно удобный и имеющий хорошо интерпретируемый результат инструмент для сравнения альтернативных вариантов и принятия решений при наличии множества критериев. Его высокая степень «настраиваемости» позволяет адаптировать метод к различным ситуациям и потребностям.

Разнообразие областей применения рассматриваемого методологического подхода, а также возможность работы с разнородными данными позволяет использовать его в новой предметной области: для комплексной оценки эффективности алгоритмов. С его помощью можно учесть различные аспекты работы алгоритмов и суммировать их в единую метрику, которая представляется более удобной в рамках анализа результата работы, описываемого множеством частных критериев, что в конечном счете значительно повышает объективность оценки.

Интегральные показатели, успешно применяемые для анализа и оценки в социально-экономических областях [11–25], ранее не были использованы для оценки алгоритмов. Новизна предлагаемого подхода заключается в адаптации методологии для использования в сфере информационных технологий. Основное внимание уделено оценке эффективности алгоритмов формирования рекомендаций, что открывает новые возможности для их анализа и улучшения их производительности.

1.2. Метрики

Для формирования интегрального показателя были выбраны основные метрики оценки алгоритмов машинного обучения и персонализированных рекомендательных систем.

• *AveragePopularity* – средняя популярность рекомендуемых объектов [26, 27]:

$$AveragePopularity = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{i\in R(u)}\phi(i)}{|R(u)|},$$

где U – множество всех пользователей, для которых сформированы рекомендации; $|U|$ – количество пользователей, для которых сформированы рекомендации; u – конкретный пользователь, данные которого используются при расчете; $R(u)$ – множество рекомендаций; $|R(u)|$ – количество элементов во множестве рекомендаций; $\phi(i)$ – количество рекомендаций элемента i в обучающих данных.

Если *AveragePopularity* принимает высокие значения, это может свидетельствовать о том, что система часто рекомендует популярные элементы, что может быть полезным для привлечения внимания пользователей. С другой стороны, слишком высокая средняя популярность может привести к недостаточному разнообразию рекомендаций.

• *GAUC* (*Grouped Area under the Curve*, сгруппированная область под кривой) – характеризует качество ранжирования для всех пользователей:

$$GAUC = \frac{1}{\sum_{u\in U}|R(u)|}\sum_{u\in U}|R(u)|\,AUC(u).$$

Здесь $AUC(u)$ – метрика оценки качества моделей в задачах классификации и ранжирования:

$$AUC(u) = \left(|R(u)|(n+1) - \frac{|R(u)|(|R(u)|+1)}{2} - \sum_{i=1}^{|R(u)|} rank_i\right) \Big/ \left(|R(u)|(n-|R(u)|)\right),$$

где $rank_i$ – нисходящий ранг i-го элемента множества $R(u)$; n – общее количество взаимодействий пользователя с элементом.

Метрика *AUC* отражает площадь под кривой *ROC* (*Receiver Operating Characteristic*), представляющей график, который показывает отношение доли правильно классифицированных положительных случаев к доле неправильно классифицированных положительных случаев при варьирова-

нии порога для принятия решения [28]. Эта метрика дает количественную оценку того, насколько модель способна проводить различия между классами (например, положительными и отрицательными случаями) при различных порогах [29].

Показатель *GAUC* демонстрирует, насколько хорошо модель ранжирует рекомендации для пользователей. Иными словами, насколько эффективно модель различает рекомендации, удовлетворяющие и не удовлетворяющие пользователя. Эта метрика используется для задач, где важно ранжирование предложений, таких как рекомендации товаров или контента.

• *GiniIndex* (индекс Джини) характеризует разнообразие распределения сформированных рекомендаций:

$$GiniIndex = \frac{\sum_{i=1}^{|I|}(2i-|I|-1)P(i)}{|I|\sum_{i=1}^{|I|}P(i)},$$

где $P(i)$ – популярность каждого элемента после обучения алгоритма; $|I|$ – рейтинг рекомендуемых объектов в неубывающем порядке.

GiniIndex используется для измерения того, насколько равномерно распределены рекомендации для каждого пользователя. Более равномерное распределение означает, что рекомендации более разнообразны и удовлетворяют различным интересам пользователя. Если распределение близко к идеальному, то *GiniIndex* будет принимать значение около нуля. Если рекомендации сосредоточены вокруг небольшого числа элементов, индекс Джини будет ближе к максимальному значению [30].

• *HitRate* (усеченный коэффициент попадания) – отношение числа рекомендаций, удовлетворяющих пользователя, к общему числу рекомендаций [31]:

$$HitRate = \frac{1}{|U|}\sum_{u\in U}|\overline{R(u)}\cap R(u)|,$$

где $\overline{R(u)}$ – множество рекомендаций, удовлетворяющих пользователя.

HitRate указывает на то, какая доля рекомендаций была действительно полезна для пользователя; чем его значение выше, тем более успешна система рекомендаций.

• *Precision* (положительное прогностическое значение) – доля релевантных элементов среди всех рекомендуемых [32]:

$$Precision = \frac{1}{|U|}\sum_{u\in U}\frac{|\overline{R(u)}\cap R(u)|}{|\overline{R(u)}|}.$$

Данный показатель демонстрирует, насколько точно система выбирает сущности, которые пользователь предпочтет. Чем выше значение *Precision*, тем более точными являются рекомендации.

• *Recall* – это мера для вычисления доли релевантных элементов из всего множества сформированных рекомендаций [33]:

$$Recall = \frac{1}{|U|}\sum_{u\in U}\frac{|\overline{R(u)}\cap R(u)|}{|R(u)|}.$$

Она указывает на то, какую долю релевантных элементов система смогла учесть в своих рекомендациях.

Метрики *Recall* и *Precision* на первый взгляд похожи, но имеют принципиальные отличия. *Recall* фокусируется на обнаружении как можно большего числа релевантных элементов и минимизации упущенных возможностей, а *Precision* – на точности выбора элементов и минимизации предоставления нерелевантных рекомендаций.

Высокое значение *Recall* означает, что система хорошо охватывает интересы пользователя, но может включать в рекомендации больше шума, а *Precision* – что система предоставляет точные рекомендации, но может упускать некоторые интересные элементы.

• Энтропия Шеннона – разнообразие рекомендаций, сформированных для пользователя [34]:

$$ShannonEntropy = -\sum_{i=1}^{|I|}p(i)\log p(i),$$

где $p(i)$ – вероятность рекомендации объекта i.

Если рекомендации разнообразны и покрывают различные интересы пользователя, энтропия будет высокой. Оценка энтропии в рекомендательных системах может быть полезной для оптимизации баланса между персонализацией (предоставление рекомендаций, соответствующих уникальным интересам пользователя) и разнообразием (предоставление рекомендаций, покрывающих более широкий спектр интересов).

• *MAP* (*Mean Average Precision*) – общее качество ранжирования элементов:

$$MAP@K = \frac{1}{|U|}\sum_{u\in U}\left(\frac{1}{\min(|\overline{R(u)}|,K)}\times\right.$$

$$\left.\times\sum_{j=1}^{|\overline{R(u)}|}|I|\overline{R(u)}_j\,Precision\right),$$

$$\overline{R(u)}_j \in R(u),$$

где K – усеченное количество сформированных рекомендаций; j – индекс рекомендации, удовлетворяющей пользователя.

Усеченное количество сформированных рекомендаций является выборкой наиболее рекомендуемых объектов и задается разработчиком системы. Оно может принимать любое значение, не превышающее количество объектов рекомендации. На практике чаще всего используется значение 10.

Данный показатель является полезной метрикой для оценки качества моделей рекомендательных систем, особенно когда важно учесть ранжирование рекомендаций. Каждый пользователь рассматривается отдельно, что позволяет учесть индивидуальные предпочтения и интересы. Поскольку *MAP* усредняет значение *Precision* всех пользователей, эта метрика предоставляет обобщенную оценку ранжирования рекомендательной системы.

• *MRR* (*Mean reciprocal rank*, средний взаимный ранг) – качество ранжирования первого элемента в списке рекомендуемых [35]:

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

где $rank_u^*$ – это ранговая позиция первого релевантного элемента, найденного алгоритмом для пользователя.

Данный показатель широко используется для оценки качества поисковых систем, рекомендательных систем и решения других задач, где важно учитывать ранжирование результатов. Чем выше *MRR*, тем лучше. Значение *MRR* будет равно единице, если релевантный элемент всегда находится на первой позиции в ранжированном списке.

• *NDCG* (*Normalized discounted cumulative gain*, нормализованный дисконтированный совокупный выигрыш) – это показатель качества ранжирования, при котором учитывается соотношение позиций и релевантности элементов в ранжированном списке [36]:

$$NDCG@K = \frac{1}{|U|}\sum_{u\in U}\left(\frac{1}{\sum_{i=1}^{\min(|R(u)|,K)}\frac{1}{\log_2(i+1)}}\times\right.$$
$$\left.\times\sum_{i=1}^{K}\delta(i\in R(u))\frac{1}{\log_2(i+1)}\right),$$

где δ – индикаторная функции (если $i\in R(u)$, то δ = 1, иначе δ = 0); i – рекомендация входящая в усеченный список.

Показатель *NDCG* широко используется для оценки качества ранжирования в рекомендательных системах, особенно в тех случаях, когда важен не только факт наличия релевантных рекомендаций, но и их порядок в списке.

Кроме того, в ходе проведения экспериментов использовались метрики, указывающие на затраченные память и время. Затраченная оперативная память вычислялась при помощи пакета memory_profiler языка Python. Под временем обучения алгоритма понимается настройка гиперпараметров или обучение слоев эмбеддинга на основе метрики *Recall*. Настройка гиперпаметров возможна на основании одной метрики. Показатель *Recall* был выбран в связи с тем, что эта характеристика важна для решения задач, в которых упущенные данные могут привести к серьезным последствиям.

Вычисления производились на компьютере с процессором Intel(R) Xeon(R) Silver 4214R @ 2.40GHz и 132 Gb RAM. При воспроизведении результатов время, замеряемое в ходе проведения эксперимента, может отличаться в зависимости от используемого вычислительного оборудования.

2. МАТЕРИАЛЫ И ИНСТРУМЕНТЫ ИССЛЕДОВАНИЯ

2.1. Наборы данных

При формировании наборов данных учитывались следующие условия. Для обеспечения разнообразия исходного материала и адекватности исследования было принято решение использовать три набора данных, в каждом из которых представлены различные объемы информации о предпочтениях пользователей. Первые два набора – MovieLens 100k и MovieLens 1m – включают в себя 100 тысяч и 1 миллион записей соответственно. Оба набора содержат ценную информацию о рейтингах, присвоенных пользователями фильмам, а также демографические данные. Наборы MovieLens используются для исследований в области рекомендательных систем и машинного обучения. В качестве третьего набора данных был выбран Amazon Gift Card, входящий в пакет Amazon Review Data. Он относится к другой предметной области нежели MovieLens и характеризуется большой разреженностью данных. Данные Amazon Review Data (2018) представляют собой набор отзывов из интернет-магазина Amazon, включающий информацию о продукте, пользователе, оценках и тексте отзыва. Этот набор содержит около 35 миллионов отзывов за 18 лет и является одним из самых используемых в машинном обучении. Набор данных Amazon Review Data (2018) может быть использован для различных целей, таких как анализ тональности отзывов, обработка естественного языка, а также для обучения моделей. Для исследований был взят тематический набор данных

Amazon Gift Card, включающий в себя информацию об отзывах на подарочные карты.

Характеристики используемых наборов данных представлены в табл. 1.

Таблица 1

Характеристики наборов данных

Параметр	Наборы данных		
	MovieLens 100k	MovieLens 1m	Amazon Gift Card
Количество пользователей	944	6041	128878
Среднее количество действий пользователей	106,04	165,59	1,1421
Количество предметов	1683	3707	1549
Среднее количество действий с предметом	59,45	269,88	95,08
Количество пересечений	100000	1000209	147194
Разреженность набора данных, %	93,70	95,50	99,92

Таким образом, подобранные наборы удовлетворяют описанным выше условиям. При этом MovieLens 1m представляет собой набор данных с бóльшим количеством записей, по сравнению с MovieLens 100k, что позволит проанализировать влияние числа записей в наборе на оценку алгоритма.

При сопоставлении Amazon Gift Card с MovieLens наблюдается существенное различие в среднем количестве действий, совершаемых пользователями. В MovieLens среднее количество действий пользователя более 100, а в Amazon Gift Card это число немногим превышает 1. Такая разница в количестве создает значительные проблемы для рекомендательных алгоритмов. При работе с Amazon Gift Card и его сравнительно ограниченным количеством действий пользователей рекомендательные алгоритмы сталкиваются с трудностями в предсказании и обобщении предпочтений. Усеченное количество данных может привести к недостаточной репрезентативности образцов, что затрудняет точные прогнозы и рекомендации.

2.2. Алгоритмы

В рамках исследования оценивались следующие рекомендательные алгоритмы:

- BPR – Bayesian Personalized Ranking from Implicit Feedback [37];
- LINE – Large-Scale Information Network Embedding [38];
- NeuCF – Neural Collaborative Filtering [39];
- DMF – Deep Matrix Factorization [40];
- SpectralCF – Spectral Collaborative Filtering [41];
- LightGCN – Simplifying and Powering Graph Convolution Network for Recommendation [42];
- MultiVAE –Variational Autoencoders for Collaborative Filtering [43];
- CDAE – Collaborative Denoising Auto-Encoders [44];
- RaCT – Ranking-Critical Training for Collaborative Filtering [45];
- SLIM – Sparse Linear Method [46];
- ItemKNN – Item-based collaborative filtering [47];
- DiffRec – Diffusion Recommender Model [48].

Данный перечень был составлен с учетом необходимости охвата различных алгоритмов формирования рекомендаций для обеспечения максимального разнообразия. Важно отметить, что в настоящее время существует гораздо большее количество подходов, чем рассмотрено в данной работе. Выбор был сделан с целью обеспечения адекватного представления разнообразных алгоритмов и их особенностей в контексте формирования рекомендаций. В то же время стоит подчеркнуть, что множество других алгоритмов также заслуживает внимания и может быть объектом дальнейших исследований в данной области.

В качестве платформы реализации алгоритмов была выбрана библиотека с открытым исходным кодом RecBole, разработанная на языке программирования Python и фреймворке машинного обучения PyTorch. Она предлагает широкий спектр алгоритмов и подходов к построению рекомендаций, а также инструменты для разработки, тестирования и оценки рекомендательных алгоритмов [49].

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

3.1. Расчет показателей работы алгоритмов

Расчеты производились для всех указанных выше наборов данных. Для наглядности промежуточные этапы описаны только для MovieLens 100k. При сборе метрик для параметра K (усеченного количества сформированных рекомендаций) использовалось значение 10.

Все собранные метрики, за исключением памяти, времени подготовки, времени предсказания и средней популярности (*Average Popularity*), представляют собой коэффициенты, принимающие значения в диапазоне от 0 до 1. Затраченная па-

мять представлена в мегабайтах, временные показатели – в секундах, средняя популярность – в количестве взаимодействий пользователей с объектом. Собранные данные отражены в табл. 2–4.

3.2. Модель

Была сформирована модель интегрального показателя: 13 параметров были «свернуты» в четыре субиндекса второго слоя; характеристики *Preparation time* (время подготовки), *Prediction time* (время формирования рекомендаций) и *Memory* (объем затраченной памяти) были агрегированы в субиндекс Resources; *Recall* и *Precision* были приведены к Accuracy; метрики *GAUC*, *MMR*, *NDCG*, *HitRate* и *MAP* были «свернуты» в субиндекс Ranking; показатели *Average Popularity*, *Gini Index* и *Shannon Entropy* обобщены в субиндекс Diversity. Структура интегрального показателя представлена на рис. 1.

Данная модель строилась исходя из принципа логического объединения параметров.

3.3. Вычисления

Для создания интегрального показателя необходимо решить три основные задачи:

• произвести нормализацию частных критериев, так как они имеют различную размерность и единицы измерений;

• произвести расчет весовых коэффициентов на слоях сети;

• определить принцип свертки частных критериев в интегральный показатель и его структурные элементы.

Для удобства построения обобщенной меры значения частных критериев должны удовлетворять следующим требованиям:

• все частные критерии должны быть безразмерны;

• для сравнения различных объектов между собой значения частных критериев должны изменяться в одном диапазоне, например, от 0 до 1;

• все частные критерии должны быть однонаправленны.

Для того чтобы исходные данные удовлетворяли всем описанным выше критериям, была проведена минимаксная нормализация значений:

$$e_{ij} = \frac{x_{ij} - \min\limits_{k=1,n}(x_{kj})}{\max\limits_{k=1,n}(x_{kj}) - \min\limits_{k=1,n}(x_{kj})},$$

где e_{ij} – нормализованное значение j-й метрики для i-го алгоритма; x_{ij} – фактическое значение j-й метрики для i-го алгоритма; n – количество алгоритмов.

Для некоторых показателей, таких как *Prediction time*, *Preparation time*, *Memory* и *Average popularity*, была применена нормализация с инверсией значений в том же числовом диапазоне от 0 до 1, так как рост значений этих критериев соответствует ухудшению оценки:

$$e_{ij} = 1 - \frac{x_{ij} - \min\limits_{k=1,n}(x_{kj})}{\max\limits_{k=1,n}(x_{kj}) - \min\limits_{k=1,n}(x_{kj})}.$$

В табл. 5 представлены нормализованные значения показателей.

Для формирования интегрального показателя предлагается энтропийный метод нахождения весовых коэффициентов частных критериев, который основан на анализе оценок среднеквадратических отклонений частных критериев каждого из них, получаемых по всей совокупности исследуемых объектов [50].

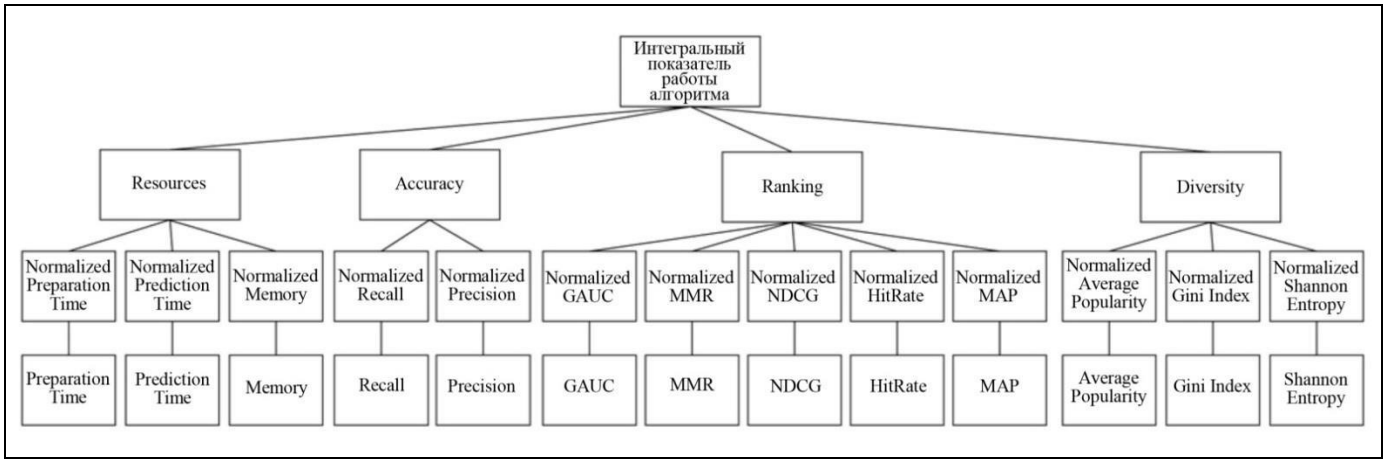

Рис. 1. Сформированный интегральный показатель

Таблица 2

Показатели работы алгоритмов с набором данных MovieLens 100k

Алгоритм	Память, мегабайт	Время подготовки (T prep.), с	Время предсказания (T pred.), с	Метрики									
				Recall	Precision	GAUC	MMR	NDCG	HitRate	MAP	Average Popularity	Gini Index	Shannon Entropy
BPR	290	20,275	0,192	0,239	0,191	0,918	0,482	0,286	0,772	0,174	241,928	0,925	0,0116
LINE	431,1	19,703	0,189	0,217	0,178	0,914	0,438	0,256	0,751	0,147	173,535	0,864	0,0104
NeuCF	454	40,751	0,923	0,238	0,189	0,919	0,459	0,277	0,766	0,165	226,602	0,911	0,0104
DMF	634,6	59,024	0,210	0,236	0,188	0,894	0,441	0,272	0,782	0,161	227,201	0,911	0,0116
SpectralCF	501,1	32,002	0,289	0,099	0,099	0,848	0,265	0,133	0,498	0,068	338,649	0,988	0,0497
LightGCN	448,3	66,221	1,191	0,226	0,177	0,909	0,453	0,268	0,750	0,161	257,067	0,945	0,0166
MultiVAE	629,6	51,722	0,532	0,237	0,182	0,911	0,463	0,276	0,766	0,165	239,964	0,923	0,0110
CDAE	516	7,853	0,220	0,098	0,096	0,698	0,279	0,135	0,504	0,071	320,434	0,986	0,0564
RaCT	404	112,653	0,662	0,263	0,201	0,918	0,486	0,298	0,808	0,178	223,918	0,892	0,0090
SLIM	416,7	0,455	0,310	0,275	0,217	0,911	0,523	0,324	0,812	0,202	246,862	0,931	0,0147
ItemKNN	521,4	0,741	0,323	0,247	0,193	0,907	0,462	0,283	0,785	0,170	217,709	0,914	0,0125
DiffRec	389,1	112,452	0,936	0,278	0,212	0,897	0,536	0,325	0,828	0,200	232,248	0,907	0,0111

Таблица 3

Показатели работы алгоритмов с набором данных MovieLens 1m

Алгоритм	Память, мегабайт	Время подготовки (T prep.), с	Время предсказания (T pred.), с	Метрики									
				Recall	Precision	GAUC	MMR	NDCG	HitRate	MAP	Average Popularity	Gini Index	Shannon Entropy
BPR	790,5	210,782	2,695	0,163	0,200	0,927	0,445	0,256	0,742	0,152	1147,6	0,885	0,0035
LINE	931,2	190,290	2,680	0,147	0,175	0,920	0,405	0,225	0,707	0,129	721,73	0,791	0,0035
NeuCF	439,5	369,691	13,162	0,145	0,184	0,924	0,403	0,229	0,710	0,132	1136,8	0,896	0,0040
DMF	926,4	1666,525	2,468	0,152	0,188	0,888	0,424	0,239	0,724	0,138	1265,9	0,928	0,0054
SpectralCF	614,7	5327,285	2,691	0,142	0,187	0,922	0,411	0,233	0,701	0,137	1142	0,891	0,0037
LightGCN	601,2	5217,235	2,815	0,155	0,166	0,924	0,390	0,214	0,743	0,112	1136,1	0,889	0,0037
MultiVAE	1019,1	409,670	4,715	0,179	0,200	0,924	0,447	0,260	0,769	0,151	1137,6	0,876	0,0034
CDAE	1215,4	517,649	3,052	0,147	0,192	0,912	0,435	0,243	0,710	0,143	1511,7	0,954	0,0055
RaCT	1122,5	556,044	4,915	0,175	0,196	0,924	0,433	0,252	0,763	0,145	1128,3	0,875	0,0033
SLIM	916,7	6,819	5,379	0,193	0,225	0,909	0,503	0,296	0,791	0,182	1421,5	0,951	0,0068
ItemKNN	1739,1	9,702	4,583	0,163	0,199	0,917	0,447	0,256	0,742	0,152	1201,2	0,918	0,0045
DiffRec	797,2	500,809	7,739	0,200	0,232	0,917	0,511	0,303	0,806	0,186	1277,6	0,922	0,0053

Таблица 4

Показатели работы алгоритмов с набором данных Amazon Gift Card

Алгоритм	Память, мегабайт	Время подготовки (T prep.), с	Время предсказания (T pred.), с	Метрики									
				Recall	Precision	GAUC	MMR	NDCG	HitRate	MAP	Average Popularity	Gini Index	Shannon Entropy
BPR	734,4	26,275	4,171	0,076	0,008	0,550	0,028	0,039	0,076	0,028	796,207	0,898	0,0115
LINE	1099,6	45,129	4,123	0,067	0,007	0,505	0,026	0,035	0,067	0,026	755,302	0,888	0,0091
NeuCF	1156,8	64,381	6,920	0,194	0,019	0,909	0,089	0,114	0,194	0,089	2974,811	0,993	0,1484
DMF	1268,3	319,145	4,464	0,219	0,022	0,862	0,093	0,122	0,219	0,093	2849,264	0,993	0,0993
SpectralCF	1356,6	179,778	3,587	0,213	0,021	0,908	0,092	0,121	0,213	0,092	2955,627	0,993	0,1501
LightGCN	1389,8	373,526	4,091	0,112	0,011	0,635	0,048	0,063	0,112	0,048	836,715	0,900	0,0087
MultiVAE	1515,4	152,994	5,120	0,315	0,032	0,915	0,125	0,170	0,315	0,125	1896,761	0,957	0,0061
CDAE	1505,8	48,741	4,803	0,212	0,021	0,786	0,093	0,121	0,212	0,093	2884,900	0,993	0,1631
RaCT	1514,7	191,619	5,590	0,319	0,032	0,913	0,125	0,170	0,319	0,125	1918,316	0,955	0,0061
SLIM	718,8	1,457	4,193	0,004	0,001	0,500	0,001	0,002	0,004	0,001	106,559	0,994	0,1921
ItemKNN	1146,1	3,545	3,960	0,172	0,017	0,741	0,073	0,096	0,172	0,073	316,197	0,875	0,0093
DiffRec	1408,8	231,541	11,727	0,231	0,023	0,758	0,101	0,132	0,232	0,101	2291,953	0,970	0,0096

Таблица 5

Нормализованные показатели работы алгоритмов для набора данных MovieLens 100k

Алгоритм	Память, мегабайт	Время подготовки (T prep.), с	Время предсказания (T pred.), с	Метрики									
				Recall	Precision	GAUC	MMR	NDCG	HitRate	MAP	Average Popularity	Gini Index	Shannon Entropy
BPR	1	0,8234	0,9965	0,7835	0,7926	0,9946	0,8007	0,8004	0,8296	0,7869	0,5858	0,4919	0,0549
LINE	0,5905	0,8285	1	0,6602	0,6786	0,9743	0,6364	0,6404	0,7653	0,5902	1	0	0,0295
NeuCF	0,5240	0,6409	0,2677	0,7790	0,7719	1	0,7172	0,7535	0,8102	0,7213	0,6786	0,3793	0,0295
DMF	0	0,4780	0,9785	0,7685	0,7669	0,8866	0,6501	0,7285	0,8584	0,6930	0,6750	0,3834	0,0549
SpectralCF	0,3874	0,7189	0,8995	0,0056	0,0248	0,6766	0	0	0	0	0	1	0,8587
LightGCN	0,5406	0,4138	0	0,7135	0,6711	0,9540	0,6932	0,7040	0,7620	0,6915	0,4941	0,6556	0,1603
MultiVAE	0,0145	0,5431	0,6576	0,7740	0,7141	0,9621	0,7309	0,7467	0,8102	0,7221	0,5977	0,4741	0,0422
CDAE	0,3442	0,9341	0,9684	0	0	0	0,0517	0,0135	0,0161	0,0179	0,1103	0,9854	1
RaCT	0,6692	0	0,5277	0,9167	0,8694	0,9955	0,8136	0,8629	0,9391	0,8204	0,6949	0,2285	0
SLIM	0,6323	1	0,8791	0,9845	1	0,9612	0,9524	0,9953	0,9518	1	0,5559	0,5397	0,1203
ItemKNN	0,3285	0,9975	0,8656	0,8290	0,8066	0,9435	0,7279	0,7858	0,8681	0,7563	0,7325	0,4003	0,0738
DiffRec	0,7124	0,0018	0,2549	1	0,9587	0,8975	1	1	1	0,9821	0,6444	0,3477	0,0443

По всем критериям оценки на каждом наборе данных было рассчитано среднеквадратическое отклонение

$$f_j = \sqrt{\frac{\sum_{i=1}^{N}(e_{ij} - \overline{e_j})^2}{N-1}},$$

где e_{ij} – нормированное значение; $\overline{e_j}$ – среднее значение метрики для всех алгоритмов; N – количество алгоритмов.

В табл. 6 представлены среднеквадратические отклонения для каждого параметра.

Далее были определены веса в каждом слое в рамках каждой группы частных критериев и субиндексов (v_j – веса в первом слое); для расчета весовых коэффициентов необходимо среднеквадратическое отклонение конкретного параметра разделить на сумму среднеквадратических отклонений в группе:

$$v_j = \frac{f_j}{\sum_{j=1}^{N_p} f_j},$$

где N_p – количество метрик в группе p.

Затем было вычислено значение каждого субиндекса p для всех алгоритмов в первом слое посредством аддитивной свертки:

$$a_{ip} = \sum_{j=1}^{N_p} e_{ij}\, v_j\ .$$

Следующим шагом необходимо вычислить среднеквадратические отклонения в субиндексах:

$$s_p = \sqrt{\frac{\sum_{i=1}^{N}(a_{ip} - \overline{a_p})^2}{N-1}},$$

Таблица 6

Среднеквадратические отклонения для набора данных MovieLens 100k

Метрика	Среднеквадратическое отклонение
Память	0,2198
T prep.	0,2730
T pred.	0,2914
Recall	0,2313
Precision	0,2196
GAUC	0,1718
MMR	0,2092
NDCG	0,2256
HitRate	0,2365
MAP	0,2229
Average Popularity	0,1827
Gini Index	0,2034
Shannon Entropy	0,2412

где $\overline{a}_p$ – среднее значение p-го субиндекса.

По аналогии с расчетом весов в первом слое происходит расчет весов во втором слое:

$$w_p = \frac{s_p}{\sum_{p=1}^{P} s_p},$$

где P – количество субиндексов.

В табл. 7 представлено распределение метрик по группам, их сумма отклонений в группе, веса первого и второго слоев.

Таблица 7

Расчет весов для набора данных MovieLens 100k

Метрика	Группа	Сумма отклонений в группе	Вес первого слоя	Вес второго слоя
Память	1	0,784	0,280	0,274
T prep.	1		0,348	
T pred.	1		0,371	
Recall	2	0,451	0,512	0,303
Precision	2		0,487	
GAUC	3	1,066	0,161	0,286
MMR	3		0,196	
NDCG	3		0,211	
HitRate	3		0,221	
MAP	3		0,209	
Average Popularity	4	0,627	0,291	0,135
Gini Index	4		0,324	
Shannon Entropy	4		0,384	

На рис. 2 схематически представлена модель интегрального показателя с весами каждого параметра для набора данных MovieLens 100k.

На рис. 3 схематически представлена модель интегрального показателя с весами для каждого параметра из набора данных MovieLens 1m.

На рис. 4 схематически представлена модель интегрального показателя с весами для каждого параметра из набора данных Amazon Gift.

Переход к субиндикаторам подразумевает расчет нормализованных значений. На рис. 2–4 этот слой сети не был отображен для простоты восприятия, но в действительности он есть, как и на рис. 1.

В табл. 8 представлен расчет значений в первом слое. Они характеризуют оценку по каждой группе критериев (затрачиваемые ресурсы, точность, качество ранжирования и разнообразие).

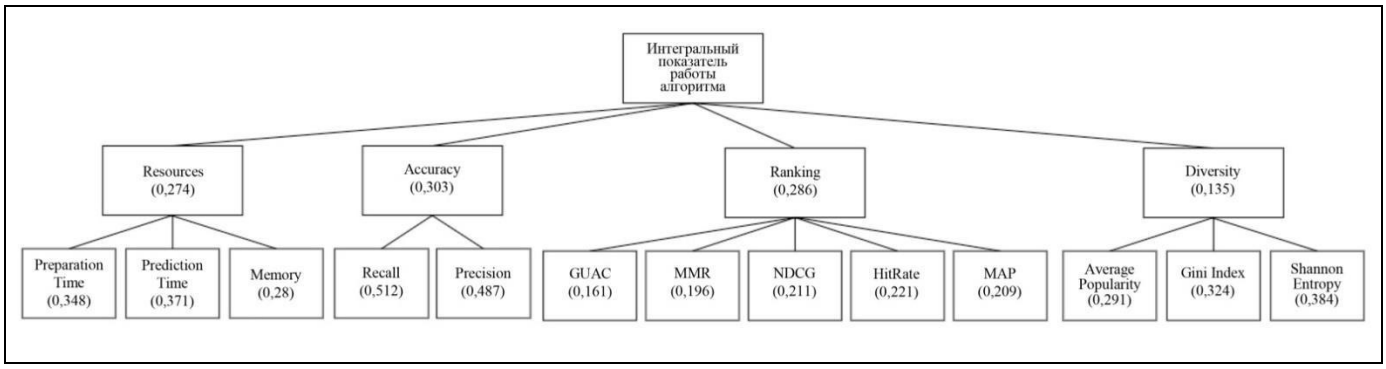

Рис. 2. Веса интегрального показателя для набора данных MovieLens 100k

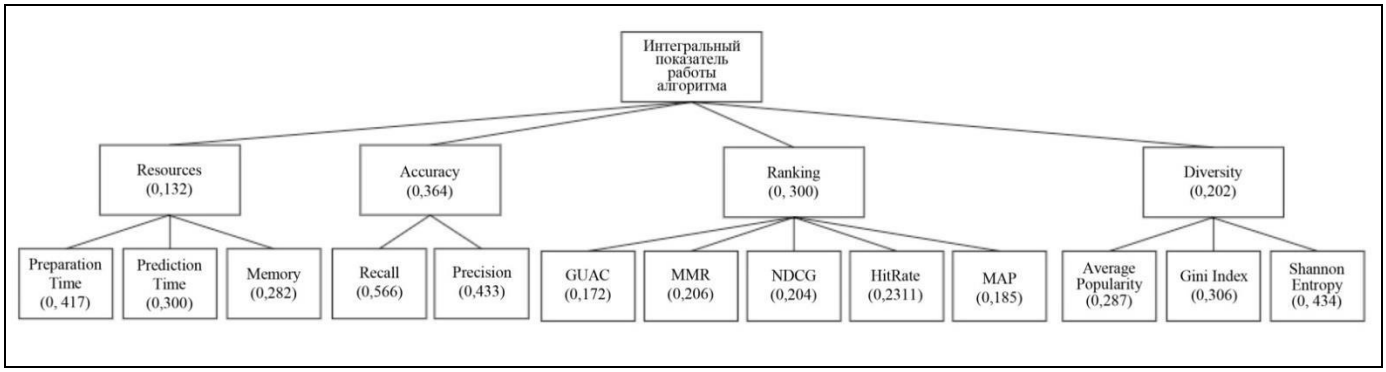

Рис. 3. Веса интегрального показателя для набора данных MovieLens 1m

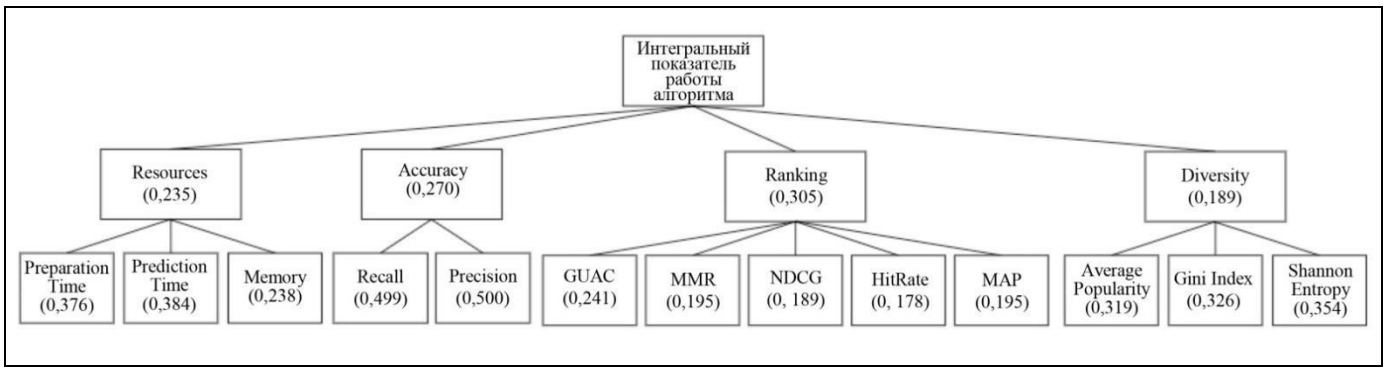

Рис. 4. Веса интегрального показателя для набора данных Amazon Gift Card

Последним шагом является реализация аддитивной свертки второго слоя по аналогии с первым:

$$b_i = \sum_{p=1}^{P} a_{ip} w_p ,$$

где P – количество значений в первом слое; i – алгоритм, для которого производится расчет.

В табл. 9 представлено итоговое значение интегрального показателя для каждого набора данных, а также среднее арифметическое значение оценки разных наборов данных. Значения в ячейках характеризуют комплексную оценку алгоритма по множеству критериев. Алгоритмы в таблице отсортированы по убыванию среднего арифметического значения оценки.

Как можно заметить из табл. 9, на примере наборов данных MovieLens 100k и 1m одни и те же алгоритмы показывают разные оценки в зависимости от размера набора. Если для первого набора данных алгоритм LINE показал высокую оценку, то для второго она упала более чем в два раза. Для алгоритма SLIM с ростом набора данных оценка практически не изменилась. Некоторые алгоритмы (DiffRec, CDAE) показали рост оценки. Из этого можно сделать вывод, что явная зависимость оценки алгоритма от объема набора данных отсутствует, это также подтверждается коэффициентом корреляции Пирсона между алгоритмами на наборах данных MovieLens 100k и 1m, значение которого составляет всего 0,56, что соответствует достаточно слабой зависимости.

Значения оценок алгоритмов на наборах данных MovieLens и Amazon Gift Card никак не соотносятся между собой. Например, алгоритм BPR, который показал высокую оценку на наборе данных MovieLens, показал низкую оценку на Amazon Gift Card.

Таблица 8

Расчет значений в первом слое для набора данных MovieLens 100k

Алгоритм	Группа / Субиндекс			
	Resources	Accuracy	Ranking	Diversity
BPR	0,9372	0,7879	0,8354	0,3512
LINE	0,8255	0,6691	0,7106	0,3026
NeuCF	0,4695	0,7755	0,7920	0,3320
DMF	0,5300	0,7677	0,7600	0,3419
SpectralCF	0,6931	0,0149	0,1091	0,6544
LightGCN	0,2956	0,6928	0,7524	0,4181
MultiVAE	0,4374	0,7448	0,7873	0,3440
CDAE	0,7814	0	0,0203	0,7361
RaCT	0,3836	0,8936	0,8826	0,2765
SLIM	0,8520	0,9920	0,9727	0,3831
ItemKNN	0,7610	0,8181	0,8120	0,3715
DiffRec	0,2950	0,9799	0,9798	0,3174

Таблица 9

Итоговая оценка

Алгоритм	MovieLens 100k	MovieLens 1m	Amazon Gift Card	Среднее арифметическое значение оценки
SLIM	0,8656	0,8390	0,4202	0,7083
DiffRec	0,7022	0,8649	0,5328	0,7000
MultiVAE	0,6184	0,5620	0,7356	0,6387
RaCT	0,6670	0,5058	0,7253	0,6327
ItemKNN	0,7402	0,4963	0,5591	0,5985
BPR	0,7834	0,5054	0,4051	0,5646
DMF	0,6426	0,3799	0,6043	0,5423
NeuCF	0,6362	0,3123	0,6525	0,5337
CDAE	0,3199	0,4090	0,6428	0,4572
LINE	0,6743	0,2874	0,3340	0,4319
SpectralCF	0,3145	0,2811	0,6506	0,4154
LightGCN	0,5637	0,2664	0,3265	0,3855

При более глубоком анализе таблицы с оценками алгоритмов можно убедиться, что первоначальная гипотеза (о том, что один и тот же алгоритм ведет себя по-разному и дает разную результативность в зависимости от набора данных) подтверждается.

Однозначно определенной причины такого явления нет, так как каждая группа алгоритмов работает на основе различных принципов. Например, алгоритмы, которые работают на основе кластеризации, могут неточно предсказывать класс объекта при недостаточности данных вследствие нечетких границ кластера. В тоже время, если данных достаточно, границы кластера полны, и кластеризация выполняется точно. Подобные проблемы могут быть и в алгоритмах, работающих на других принципах.

Также необходима настройка гиперпараметров, алгоритмов в соответствии с контекстом данных, учитывающий больше одной характеризующей метрики. Как раз для решения этой задачи подходит единая метрика.

4. ДИСКУССИЯ

Результаты исследования свидетельствуют о возможности практического применения предложенного подхода для подбора оптимального рекомендательного алгоритма, особенно в контексте конкретных наборов данных. Важно отметить, что дальнейшее усовершенствование метода может быть достигнуто путем введения пользовательского ранжирования. Это даст возможность учитывать индивидуальные предпочтения пользователей, отдавая приоритет тем критериям, которые для них наиболее важны. Например, пользователь, который ценит точность рекомендаций больше, чем скорость работы алгоритма, может выставить соответствующие веса, делая подход более гибким и персонализированным.

Также можно отметить потенциал применения данного подхода для оптимизации гиперпараметров алгоритмов. Это позволяет систематически оценивать влияние изменения гиперпараметров на конкретные метрики и общую оценку алгоритма. Систематическое оценивание гиперпараметров с помощью данного подхода позволяет создавать итеративный процесс улучшения моделей. Разработчики смогут регулярно анализировать и оптимизировать параметры в соответствии с изменяющимися требованиями и данными.

Оптимизация гиперпараметров позволяет улучшить производительность алгоритма и повысить его точность. Она также может помочь в понимании того, какие параметры наиболее важны для достижения наилучших результатов. Это важно, поскольку различные задачи могут требовать различных настроек параметров для достижения оптимальных результатов.

Важным направлением развития метода может быть его применение для формирования ансамблей алгоритмов рекомендательных систем. Максимизация оценки позволяет эффективно комбинировать различные алгоритмы в ансамбль с учетом их вклада в общую рекомендательную систему. Такой подход позволяет создать более надежную и эффективную систему рекомендаций, которая учитывает сильные и слабые стороны каждого алгоритма. Это позволяет сгладить недостатки отдельных алгоритмов и создать более точные и релевантные рекомендации для пользователей.

В представленных выше подходах рассматривались исключительно офлайн-оценки рекомендательных алгоритмов, что, однако, может сформировать неполное представление об их эффективности. В перспективе может быть разработан интегральный показатель, который обеспечит онлайн-оценку работы таких алгоритмов. Возможен также сценарий объединения онлайн- и офлайн-метрик в обобщенную меру. В основе этих параметров могут лежать такие показатели, как *Gross merchandise volume* (GMV), *Click-through rate* (CTR), *Conversion rate* (CVR). Эти метрики позволят целостно оценить влияние рекомендательных алгоритмов на бизнес-показатели, учитывая как офлайн-, так и онлайн-аспекты их деятельности.

Дополнительные направления исследований могут включать в себя анализ влияния динамических изменений в данных на производительность рекомендательных алгоритмов. Они могут включать в себя изучение эффективности метода в условиях изменяющихся предпочтений пользователей, сезонных колебаний или временны́х трендов. Исследование влияния факторов, таких как смена трендов в потребительском поведении, может быть ключевым шагом к разработке более адаптивных и устойчивых рекомендательных систем.

ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена задача формирования интегрального показателя для оценки эффективности рекомендательных систем. Для этого были выбраны различные алгоритмы и наборы данных, включая MovieLens и Amazon Gift Card. Для каждого алгоритма и набора данных были рассчитаны различные метрики, которые пред-

ставляют собой различные аспекты качества рекомендаций, отражают их точность, полноту, разнообразие и другие характеристики. Интегральный показатель был создан путем объединения этих метрик с помощью энтропийного метода расчета весов. Этот подход позволяет учитывать относительную важность каждой метрики и обеспечивать сбалансированность в интегральной оценке эффективности рекомендательных систем.

Применение такого подхода представляется новым в области оценки рекомендательных систем. Интегральный показатель, созданный с использованием энтропийного метода, представляет собой уникальный инструмент для сравнения различных алгоритмов и наборов данных с точки зрения их общей производительности. Этот подход позволяет учитывать разнообразные аспекты качества рекомендаций и делает его применимым в различных сценариях использования рекомендательных систем.

Результаты показали, что интегральный показатель может быть полезным инструментом для оценки эффективности рекомендательных систем, обладающим способностью объединять разнообразные метрики в единую общую оценку, что придает ему превосходство в информативности и удобстве использования по сравнению с отдельными критериями. Интегральный показатель предоставляет более глубокое и всеобъемлющее понимание эффективности рекомендательных систем, что делает его ценным инструментом для принятия обоснованных решений в данной области.

Однако стоит отметить, что создание интегрального показателя требует проведения большого объема вычислений и анализа данных. Кроме того, выбор алгоритмов и наборов данных может существенно влиять на результаты оценки. Поэтому для получения более точных результатов необходимо проводить дополнительные исследования и эксперименты с различными алгоритмами и наборами данных. Эти исследования позволят лучше понять влияние различных параметров на конечные результаты и оптимизировать процесс создания интегрального показателя. Такой подход позволит оценить, насколько обобщенная мера способна адаптироваться к различным сценариям использования, и выявить ее ограничения.

Дополнительные исследования также могут помочь оптимизировать весовые коэффициенты, используемые при объединении различных метрик в интегральный показатель. Это важно для того, чтобы учесть относительную важность каждой метрики и обеспечить баланс между различными аспектами качества рекомендаций.

Таким образом, дальнейшие исследования в области создания интегральных показателей для оценки эффективности рекомендательных систем являются необходимыми для повышения их точности, обобщенности и практической применимости.

ЛИТЕРАТУРА

1. *Fayyaz, Z., Ebrahimian, M., Nawara, D.,* et al. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities // Appl. Sci. – 2020. – Vol. 10, no. 21. – Art. no. 7748. – DOI: https://doi.org/10.3390/app10217748.
2. *Amatriain, X., Basilico, J.* Recommender Systems in Industry: A Netflix Case Study. Recommender Systems Handbook / Ed. by F. Ricci, L. Rokach, B. Shapira. – Boston, MA: Springer, 2015. – P. 385–419. – DOI: https://doi.org/10.1007/978-1-4899-7637-6_11.
3. *Wibisono, C., Purwanti, E., Effendy, F.* A Systematic Literature Review of Movie Recommender Systems for Movie Streaming Service // AIP Conf. Proc. – 2023. – Vol. 2554, no. 1. – Art. no. 040005. – DOI: https://doi.org/10.1063/5.0104316.
4. *Kulshin, R., Sidorov, A., Senchenko, P.* Using Neural Networks with Reinforcement in the Tasks of Forming User Recommendations // Journal of Physics: Conference Series. – 2022. – Vol. 2291, no. 1. – Art. no. 012005. – DOI: 10.1088/1742-6596/2291/1/012005.
5. *Sipser, M.* Introduction to the Theory of Computation. Course Technology. – Boston, MA: PWS Publishing Co., 2006. – 227 p.
6. *Aixin, S.* On Challenges of Evaluating Recommender Systems in an Offline Setting // Proceedings of the 17th ACM Conference on Recommender Systems. – Singapore, 2023. – P. 1284–1285.
7. *Jimenez-Fernandez, E., Ruiz-Martos, M.* Review of some statistical methods for constructing composite indicators // Studies of Applied Economics. – 2020. – Vol. 38, no. 1. – p. 1–15. – DOI: https://doi.org/10.25115/eea.v38i1.3002.
8. *Nardo, M., Saisana, M., Saltelli, A., Tarantola, S.* Tools for Composite Indicators Building. Report no. JRC31473. – Rome, Italy: European Commission, ISPRA, 2005.
9. *Becker, W., Saisana, M., Paruolo, P., Vandecasteele, I.* Weights and Importance in Composite Indicators: Closing the Gap. Ecological Indicators. – 2017. – Vol. 80. – P. 12–22. – DOI: https://doi.org/10.1016/j.ecolind.2017.03.056.
10. *Сидоров А.А.* Методологический подход к интегральной оценке состояния и динамики многомерных объектов социально-экономической природы // Проблемы управления. – 2016. – № 3. – С. 32–40. [*Sidorov, A.A.* Methodological Approach to the Integral Assessment of the State and Dynamics of Multidimensional Objects of Socio-Economic Nature / Control Sciences. – 2016. – No. 3 – P. 32–40. (In Russian)].
11. *Abberger, K., Graff, M., Müller, O., Sturm, J.* Composite Global Indicators from Survey Data: The Global Economic Barometers // Rev. World Econ. – 2022. – Vol. 158. – P. 917–945. – DOI: https://doi.org/10.1007/s10290-021-00449-8.
12. *Endrodi-Kovacs, V., Tankovsky, O.* A Composite Indicator for Economic Integration Maturity: The Case of Western Balkan Countries // Eastern Journal of European Studies. – 2022. – Vol. 13, no 1. – P. 148–166. – DOI: 10.47743/ejes-2022-0107.

13. *Khadzhynova, O., Simanaviciene, Z., Mints, O.,* et al. Assessment of the EU Countries' Economic Security Based on the Composite Indicators // WSEAS Transactions on Business and Economics. – 2022. – Vol. 19. – P. 690–700. – DOI: 10.37394/23207.2022.19.61.
14. *McDonnell, T., Cosgrove, G., Hogan, E.,* et al. Methods to Derive Composite Indicators Used for Quality and Safety Measurement and Monitoring in Healthcare: A Scoping Review Protocol // BMJ Open. – 2023. – Vol. 13, no. 7. – DOI: 10.1136/bmjopen-2022-071382.
15. *Kara, P., Valentin, J., Mainz, J., Johnsen, S.* Composite Measures of Quality of Health Care: Evidence Mapping of Methodology and Reporting // PLoS One. – 2022. – Vol. 17, no. 5. – DOI: 10.1371/journal.pone.0268320.
16. *Asadi-Lari, M., Majdzadeh, R., Mansournia, M.A,* et al. Construction and Validation of CAPSES Scale as a Composite Indicator of SES for Health Research: An Application to Modeling Social Determinants of Cardiovascular Diseases // BMC Public Health. – 2023. – Vol. 23. – Art. no. 293. – DOI: https://doi.org/10.1186/s12889-023-15206-9.
17. *Abenayake, C., Mikami, Y., Matsuda, Y., Jayasinghe, A*. Ecosystem Services-Based Composite Indicator for Assessing Community Resilience to Floods // Environmental Development. – 2018. – Vol. 27. – P. 34–46. – DOI: https://doi.org/10.1016/j.envdev.2018.08.002.
18. *Gómez-Limón, J., Arriaza, M., Guerrero-Baena, M*. Building a Composite Indicator to Measure Environmental Sustainability Using Alternative Weighting Methods // Sustainability. – 2020. – Vol. 12, no. 11. – Art. no. 4398. – DOI: https://doi.org/10.3390/su12114398.
19. *Alam, M., Dupras, J., Messier, C.* A Framework towards a Composite Indicator for Urban Ecosystem Services // Ecological Indicators. – 2016. – Vol. 60. – P. 38–44. – DOI: https://doi.org/10.1016/j.ecolind.2015.05.035.
20. *Melo-Aguilar, C., Agulles, M., Jordà, G.* Introducing Uncertainties in Composite Indicators. The Case of the Impact Chain Risk Assessment Framework // Front. Clim. – 2022. – Vol. 4. – DOI: https://doi.org/10.3389/fclim.2022.1019888.
21. *Dolge, K., Blumberga, D.* Composite Risk Index for Designing Smart Climate and Energy Policies. Environmental and Sustainability Indicators. – 2021. – Vol. 12. – Art. no. 100159. – DOI: https://doi.org/10.1016/j.indic.2021.100159.
22. *Do, H., Ly, T., Do, T.* Combining Semi-quantitative Risk Assessment, Composite Indicator and Fuzzy Logic for Evaluation of Hazardous Chemical Accidents // Sci. Rep. – 2020. – Vol. 10. – Art. no. 18544. – DOI: https://doi.org/10.1038/s41598-020-75583-8.
23. *Avanesian, G., Mizunoya, S., Delamonica, E*. UNICEF Remote Learning Readiness Index: A Composite Indicator to Assess Resilience of Education Sector against Crises and Emergencies // Statistical Journal of the IAOS. – 2022. – Vol. 38. – P. 1–14. – DOI: 10.3233/SJI-220051.
24. *Segovia-Gonzalez, M., Contreras, I.* A Composite Indicator to Compare the Performance of Male and Female Students in Educational Systems // Soc. Indic. Res. – 2023. – Vol. 165. – P. 181–212. – DOI: https://doi.org/10.1007/s11205-022-03009-1.
25. *Hubelova, D., Odvarkova, V., Chalupa, P*. Selected Factors of Education Level in East African Countries: Comparative Method Using Composite Indicator // Geographical Journal. – 2016. – Vol. 68. – P. 55–72.
26. *Silveira, T., Zhang, M., Lin, X.,* et al. How Good Your Recommender System Is? A Survey on Evaluations in Recommendation. Journal of Machine Learning and Cybernetics. – 2016. – Vol. 10. – P. 813–831. – DOI: https://doi.org/10.1007/s13042-017-0762-9.
27. *Hongzhi, Y., Cui, B., Li, J.,* et al. Challenging the Long Tail Recommendation // Proceedings of the VLDB Endowment. – 2012. – Vol. 5. – P. 896–907. – DOI: https://doi.org/10.14778/2311906.2311916.
28. *Hanczar, B., Hua, J., Sima, C.,* et al. Small-Sample Precision of ROC-Related Estimates // Bioinformatics. – 2010. – Vol. 26. – P. 822–830. – DOI: https://doi.org/10.1093/bioinformatics/btq037.
29. *Calders, T., Jaroszewicz, S*. Efficient AUC Optimization for Classification. Lecture Notes in Computer Science. – 2007. – Vol. 4702. – P. 42–53. – DOI: https://doi.org/10.1007/978-3-540-74976-9_8.
30. *Wenlong, S., Khenissi, S., Nasraoui, O., Shafto, P*. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering // Proceedings of the 2019 World Wide Web Conference. – San Francisco, 2019. – P. 645–651. – DOI: https://doi.org/10.1145/3308560.3317303.
31. *Zhang, Q., Cao, L., Zhu, C.,* et al. CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. – Stockholm, 2018. – P. 3662–3668. – DOI: https://doi.org/10.24963/ijcai.2018/509.
32. *Bellogin, A., Castells, P., Cantador, I.* Precision-Oriented Evaluation of Recommender Systems: An Algorithmic Comparison // Proceedings of the Fifth ACM Conference on Recommender Systems. – Chicago, 2011. – P. 333–336. – DOI: https://doi.org/10.1145/2043932.2043996.
33. *Wang, Y*. Application of Recall Methods in Recommendation Systems // Proceedings of the 3rd International Conference on Signal Processing and Machine Learning. – Oxford, 2023. – Vol. 4. – P. 44–51. – DOI 10.54254/2755-2721/4/20230344.
34. *Zriaa, R., Sadiki, H., Ertel, M.,* et al. Qualitative Recommender System Using Entropy-Weighted Pedagogical Criteria for Effective Training in E-learning Platforms // Journal of Theoretical and Applied Information Technology. – 2023. – Vol. 101, no. 9. – P. 3517–3529.
35. *Kumar, C., Kumar, M.* User Session Interaction-Based Recommendation System Using Various Machine Learning Techniques // Multimed. Tools Appl. – 2023. – Vol. 82. – P. 21 279–21 309.
36. *Wang, Y., Wang, L, Li, Y.,* et al. A Theoretical Analysis of NDCG Ranking Measures // Proceedings of the 26th Annual Conference on Learning Theory. – Princeton, 2013. – Vol. 30. – P. 25–54.
37. *Steffen, R., Freudenthaler, C., Gantner, Z., Schmidt, L*. BPR: Bayesian Personalized Ranking from Implicit Feedback // Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. – Montreal, 2009. – P. 452–461.
38. *Jian, T., Qu, M., Wang, M.,* et al. LINE: Large-Scale Information Network Embedding // Proceedings of the 24th International Conference on World Wide Web. – Florence, 2015. – P. 1067–1077. – DOI: https://doi.org/10.1145/2736277.2741093.
39. *He, X., Liao, L., Zhang, H.,* et al. Neural Collaborative Filtering // Proceedings of the 26th International Conference on World Wide Web. – Perth, 2017. – P. 173–182. – DOI: https://doi.org/10.1145/3038912.3052569.
40. *Xue, J, Dai, X., Zhang, J.,* et al. Deep Matrix Factorization Models for Recommender Systems // Proceedings of the Twenty-Sixth International Joint Conference on Artificial In-

telligence. – Toronto, 2017. – P. 3203–3209. – DOI: https://doi.org/10.24963/ijcai.2017/447.
41. *Lei, Z., Lu, C., Jiang, F.,* et al. Spectral Collaborative Filtering // Proceedings of the 12th ACM Conference on Recommender Systems. – Vancouver, 2018. – P. 311–319. – DOI: https://doi.org/10.1145/3240323.3240343.
42. *He, X., Deng, K., Wang, X.,* et al. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. – Xian, 2020. – P. 639–648. – DOI: https://doi.org/10.1145/3397271.3401063.
43. *Dawen, L., Krishnan, R., Hoffman, M., Jebara, T*. Variational Autoencoders for Collaborative Filtering // Proceedings of the 2018 World Wide Web Conference. – Lyon, 2018. – P. 689–698. – DOI: https://doi.org/10.1145/3178876.3186150.
44. *Yao, W., DuBois, C., Zheng, A., Ester, M*. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems // Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. – San Francisco, 2016. – P. 153–162. – DOI: https://doi.org/10.1145/2835776.2835837.
45. *Lobel, S., Li, C., Gao, J., Carin, L*. RaCT: Toward Amortized Ranking-Critical Training for Collaborative Filtering // Proceedings of the Eighth International Conference on Learning Representations (ICLR). – Addis Ababa, 2020.
46. *Ning, X., Karypis, G*. SLIM: Sparse Linear Methods for Top-N Recommender Systems // Proceedings of the IEEE 11th International Conference on Data Mining. – Vancouver, 2011. – P. 497–506. – DOI: 10.1109/ICDM.2011.134.
47. *Mukund, D., Karypis, G*. Item-Based Top-N Recommendation Algorithms // ACM Transactions on Information Systems. – 2004. – Vol. 22. – P. 143–177. – DOI: https://doi.org/10.1145/963770.963776.
48. *Wang, W., Xu, Y., Feng, F.,* et al. Diffusion Recommender Model // Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. – Taipei, 2023. – P. 832–841. – DOI: https://doi.org/10.1145/3539618.3591663.
49. *Zhao, W., Mu, S., Hou, Y.,* et al. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms // Proceedings of the 30th ACM International Conference on Information & Knowledge Management. – Queensland, 2021. – P. 4653–4664. – DOI: https://doi.org/10.1145/3459637.3482016.
50. *Раев А.Г.* Об одном способе определения весовых коэффициентов частных критериев при построении аддитивного интегрального критерия // Автоматика и телемеханика. – 1984. – Вып. 5. – С. 162–165. [*Raev, A.G.* On One Way to Determine Weighting Coefficients of Particular Criteria in Development of an Integral Additive Criterion // Avtomatika i telemekhanika. – 1984. – No. 5. – P. 162–165. (In Russian)]

Статья представлена к публикации членом редколлегии академиком РАН Д. А. Новиковым.

Поступила в редакцию 05.03.2024,
после доработки 14.05.2024.
Принята к публикации 19.07.2024.

Кульшин Роман Сергеевич – аспирант,
✉ roman.s.kulshin@tusur.ru
ORCID ID: 0000-0002-6891-1869

Сидоров Анатолий Анатольевич – канд. экон. наук,
✉ anatolii.a.sidorov@tusur.ru
ORCID ID: 0000-0002-9236-3639

Томский государственный университет систем управления и радиоэлектроники, г. Томск

AN ENTROPY-BASED COMPOSITE INDICATOR FOR EVALUATING THE EFFECTIVENESS OF RECOMMENDER SYSTEM ALGORITHMS

R. S. Kulshin* and A. A. Sidorov**

Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia

*✉ roman.s.kulshin@tusur.ru, **✉ anatolii.a.sidorov@tusur.ru

Abstract. The problem of forming a composite indicator for evaluating the effectiveness of recommender system algorithms is considered. A novel composite indicator is proposed by combining individual metrics using the entropy method. The testing base of this study consists of 12 algorithms (on the one hand) and 3 datasets (on the other). For each algorithm–dataset combination, we calculate partial criteria used in evaluating recommender systems. According to the results presented below, the composite indicator is an effective tool for evaluating the performance of recommender system algorithms. As is shown, the performance of the algorithms varies depending on the size and other basic characteristics of a particular dataset. This indicator can be used to develop more efficient algorithms and their ensembles as well as to optimize hyperparameters and improve the quality of recommendations.

Keywords: recommender systems, composite indicator, algorithms, metrics, datasets.

Acknowledgments. This work was carried out within the state order of the Ministry of Science and Higher Education of the Russian Federation, project no. FEWM-2023-0013.