УДК 004.8

К.Э. Петров[1], И.В. Кобзев[2]

[1]ХНУРЭ, г. Харьков, Украина, kostiantyn.petrov@nure.ua

[2]ХарРИ НАГУ при Президенте Украины, г. Харьков, Украина, ikobzev12@gmail.com

ПРОГНОЗИРОВАНИЕ ПРЕДПОЧТЕНИЙ ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ АНАЛИЗА ИХ ДЕЙСТВИЙ

Предложен подход к определению предпочтений пользователей, который базируется на синтезированной модели выбора. Решена задача структурной и параметрической идентификации этой модели на основе идей теории компараторной идентификации. Данный подход дает возможность прогнозировать оценки пользователей для объектов определенной категории, что позволяет повысить релевантность выдачи рекомендаций. Приведены результаты численного моделирования, подтверждающие эффективность описанного подхода.

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ, МОДЕЛЬ ВЫБОРА, МЕТОД КОМПАРАТОРНОЙ ИДЕНТИФИКАЦИИ, ФУНКЦИЯ ПОЛЕЗНОСТИ

К.Е. Петров, І.В. Кобзев. Прогнозування переваг користувачів на базі аналізу їх дій. Запропоновано підхід до визначення переваг користувачів, який базується на синтезованої моделі вибору. Розв'язано задачу структурної та параметричної ідентифікації цієї моделі на основі ідей теорії компараторної ідентифікації. Цей підхід дає можливість прогнозувати оцінки користувачів для об'єктів певної категорії, що дозволяє підвищити релевантність видачі рекомендацій. Наведено результати чисельного моделювання, що підтверджують ефективність описаного підходу.

РЕКОМЕНДАЦІЙНІ СИСТЕМИ, МОДЕЛЬ ВИБОРУ, МЕТОД КОМПАРАТОРНОЇ ІДЕНТИФІКАЦІЇ, ФУНКЦІЯ КОРИСНОСТІ

K.E. Petrov, I.V. Kobzev. Forecasting preferences of users based on the analysis of their actions. In the article suggests an approach to determining user's preferences, which is based on the synthesized model of choice. The problem of structural and parametric identification of this model is solved on the basis of the ideas of the theory of comparative identification. This approach give opportunity it possible to predict user's estimates for objects of a certain category, which makes it possible to increase the relevance of issuing recommendations. The results of numerical modeling confirming the effectiveness of the described approach are presented.

RECOMMENDER SYSTEMS, MODEL OF SELECTION, METHOD OF COMPARATIVE IDENTIFICATION, UTILITY FUNCTION

Введение

Рекомендательные системы настолько прочно вошли в нашу жизнь, что мы зачастую даже не осознаем на сколько велико их влияние на наш выбор. Они активно используются в электронной коммерции, поисковых системах, социальных сетях, новостных сервисах, он-лайн кинотеатрах и библиотеках, т. е. фактически везде для персонализации контента, таргетирования рекламы, маркетинга и т. п.

Рекомендательные системы — это большой класс моделей и алгоритмов, относящихся к машинному обучению, которые пытаются предсказать какие объекты (товары, контент) будут интересны пользователю, на основе имеющейся определенной информации о его профиле (личных данных, интересах, истории просмотров, оценок и т. п.). Фактически эти системы прогнозируют предпочтения пользователей.

Для адекватной работы рекомендательных систем необходимо каким-то образом собирать данные о пользователях. Для этого используют сочетание явных (например, создание личного кабинета, запрос у пользователя оценки объекта по некоторой шкале и т. п.) и неявных (например, слежение за тем, что просматривает пользователь в интернет-магазине, что положил в корзину или купил) методов.

Можно условно выделить четыре типа рекомендательных систем [1].

1. Коллаборативная фильтрация (collaborative filtering) [2, 3]. Рекомендации основываются на поведенческих характеристиках (оценках) одного человека или группы людей, которые похожи между собой. Например, пользователю, со схожими с вашими оценками, понравился данный фильм, вероятно вам он тоже понравится. Этот тип систем обладает теоретически высокой точностью, однако релевантные рекомендации возможны только в том случае если известна хоть какая-то информация об интересах пользователя.

2. Основанные на контенте (content-based) [2, 3]. Здесь выдача рекомендаций осуществляется на основе данных, собранных о каждом конкретном объекте. Пользователю рекомендуются объекты, похожие на те, которыми он ранее интересовался (например, книги того же жанра или автора). К достоинствам таких систем можно отнести возможность рекомендовать даже те объекты, которые не были оценены другими пользователями или давать рекомендации новым пользователям, тем самым вовлекая их в сервис. Для этого не нужно долго

собирать данные о их предпочтениях. Основные недостатки — зависимость от предметной области, снижение точности, а также увеличение времени разработки таких систем.

3. Основанные на знаниях (knowledge-based) о предметной области [2, 3], а не о каждом объекте. Эти дополнительные знания позволяют выдавать рекомендации основываясь не просто на «похожести» чего-либо, а с более сложными условиями. Например, к этому телефону вам возможно понадобиться чехол или защитное стекло. Такой тип рекомендаций имеет высокую точность, предлагая пользователю то, что ему нужно. Однако минусом таких систем является высокая сложность разработки и сбора данных, в частности из-за необходимости изучения и анализа взаимосвязей между объектами.

4. Гибридные (hybrid) рекомендательные системы [1, 2, 3] основаны на комбинировании коллаборативных и контентных подходов, что позволяет избежать большинства недостатков, которые проявляются при использовании каждого из методов по отдельности. Например, самая технологичная на текущий момент система Netflix (BellKor) является комбинацией 27 рекомендательных алгоритмов. Главным минусом таких систем является самая высокая сложность их разработки.

Из анализа литературных источников [1, 2, 3] и многих других можно сделать вывод, что все реально работающие рекомендательные системы являются гибридными.

Общими проблемами, которые присущи всем типам рассмотренных выше систем являются: во-первых, так называемая проблема «холодного старта», когда новым или нетипичным пользователям (которых нельзя отнести ни к одной определенной группе) сложно дать рекомендации и когда новые объекты (которые еще никем не оценены) не попадают в списки рекомендаций; во-вторых, частая банальность, обыденность рекомендаций, что снижает доверие пользователей к целесообразности следовать им.

Системы, которые способны выдавать релевантные рекомендации сокращают время, необходимое для поиска различных объектов и услуг, и значительно увеличивают вероятность попадания в поле зрения пользователя других объектов, которые смогут его заинтересовать. И как результат повышают лояльность и удовлетворенность пользователей веб-сервисами. Это в свою очередь приводит к увеличению потребления и росту прибыли, а также увеличивает частоту посещений веб-ресурсов постоянными пользователями и уменьшает отток клиентов. Поэтому разработка новых методов и алгоритмов повышающих эффективность использования такого рода систем является весьма актуальной.

1. Постановка задачи

Основной проблемой выдачи пользователю релевантных рекомендаций является отсутствие информации о нем. Ведь не секрет, что получение информации о пользователе на основе явных методов является крайне проблематичным. Пользователь с большим нежеланием проходит регистрацию на сайте, выставляет оценки, формулирует мнения об объектах, заполняет анкеты и т. п. Неявные же методы позволяют получить очень поверхностную информацию. Например, по IP-адресу можно получить информацию о его местоположении, а по просмотрам категорий товара или контента — о его поле и приблизительном возрасте и т. д. В этом случае применение, например, метода коллаборативной фильтрации крайне затруднено из-за малого объема информации о пользователе.

Без потери общности рассмотрим подход к получению информации о предпочтениях пользователя в рамках анализа его поведения при выборе товара в интернет-магазине. Ведь зная его предпочтения и интересы мы можем существенно повысить релевантность рекомендаций.

Пусть потребителю (новому пользователю) предлагается некоторое ограниченное множество товаров различных марок одинакового функционального назначения (например, мобильные телефоны) $X=\{x_1,x_2,...,x_n\}$. Каждый товар из этой определенной категории может быть описан набором разнородных частных характеристик $K(x_i)=\langle k_1(x_i),k_2(x_i),...,k_m(x_i)\rangle$, $i=\overline{1,n}$. Эти характеристики достаточно полно отражают качество, надежность, функциональные свойства, безопасность, экономичность и т. п. предлагаемого товара и, кроме того, допускают их объективное количественное измерение.

Необходимо определить предпочтения потребителя на основе анализа только информации о его действиях на сайте конкретного интернет-магазина и базируясь на этих данных построить модель выбора потребителем товаров данной категории и идентифицировать её параметры.

При этом под моделью выбора будем подразумевать математическую модель, определяющую его выбор той или иной марки товара из рассматриваемой категории X.

Используя эту модель можно будет предсказать оценки товаров, которые пользователь не рассматривал (или новых товаров) и выдавать ему конкретные рекомендации.

2. Структурная и параметрическая идентификация математической модели выбора

Согласно теории поведения потребителей [3], каждой марке товара $x_i \in X$, $i=\overline{1,n}$ можно сопоставить некоторую обобщенную оценку "полезности" для потребителя. В общем виде такую обобщенную оценку, формально, можно выразить

в виде некоторой функции полезности $P(x_i)$, которая зависит от частных характеристик товара $K(x_i)$ следующим образом:

$$P(x_i) = F\left[W, K(x_i)\right],\ i = \overline{1,n}, \qquad (1)$$

где $W = \langle w_1, w_2, ..., w_s \rangle$ – кортеж параметров (коэффициенты относительной важности частных характеристик и их комплексов).

Предположим, что потребитель выбирает "самый полезный" товар из данной категории не объясняя причин своего выбора. Таким образом, потребитель из всех возможных альтернатив $x_i \in X$, $i = \overline{1,n}$ выбирает товар $x^0 \in X$ с максимальным значением функции полезности, т. е.

$$x^0 = \arg\max_{x_i \in X} P(x_i). \qquad (2)$$

На основе только этой информации необходимо определить важность для потребителя тех или иных частных характеристик товара и их комплексов (т. е. их относительных "весовых коэффициентов" W).

Функция полезности (1) и критерий выбора (2) полностью определяют математическую модель выбора.

Следующим этапом является решение задачи ее структурной (выбор и обоснование вида функции $F[...]$) и параметрической (определение значений ее параметров W) идентификации.

Как показано в работе [4] в качестве функции $F[...]$ целесообразно выбрать полином Колмогорова-Габора или некоторый его фрагмент. Главным его достоинством является, то что с его помощью можно реализовать как простейшие аддитивные и мультипликативные принципы формирования функций обобщенной полезности, так и более сложные линейные и нелинейные аддитивно-мультипликативные. Это бывает необходимо в случае рассмотрения не только отдельных частных характеристик альтернатив, но и их комплексов.

В нашем случае, будем в качестве функции полезности $P(x_i)$ будем использовать фрагмент полинома, который представляет собой аддитивную функцию. Это является оправданным из-за небольшого объема информации полученной в ходе наблюдения за поведением потребителя при выборе товара. Однако это не исключает использования более сложных моделей [4] с членами второго, третьего и более высоких порядков, учитывающих взаимовлияние частных характеристик, в случае, если удается получить больше информации от покупателей (при использовании явных методов).

Таким образом, модель обобщенной многофакторной оценки $P(x_i)$ некоторой марки товара $x_i \in X$, $i = \overline{1,n}$ можно записать в виде:

$$P(x_i) = \sum_{j=1}^{m} w_j k_j^H(x_i), \qquad (3)$$

где $K^H(x_i) = \left\langle k_1^H(x_i), k_2^H(x_i), ..., k_m^H(x_i) \right\rangle$ – нормированные значения частных характеристик товара; w_j – безразмерные коэффициенты относительной важности нормированных частных характеристик $k_j^H(x_i)$, которые удовлетворяют условиям

$$w_j \in [0,1],\ j = \overline{1,m};\ \sum_{j=1}^{m} w_j = 1. \qquad (4)$$

Необходимость нормирования частных характеристик обусловлена тем, что в общем случае, они имеют различные размерность, интервал изменений и направление доминирования. Это нормирование производится следующим образом:

$$k_j^H(x_i) = \frac{k_j(x_i) - k_j^-(x_i)}{k_j^+(x_i) - k_j^-(x_i)},\ j = \overline{1,m},\ i = \overline{1,n}, \qquad (5)$$

где $k_j(x_i)$ — действительное (абсолютное), $k_j^-(x_i)$ — "наихудшее", $k_j^+(x_i)$ –"наилучшее" значения j-й частной характеристики.

Рассмотрим этап параметрической идентификацию модели формирования многофакторной скалярной оценки "полезности" товара (3).

На этом этапе нам необходимо формализовать информацию, полученную в ходе наблюдения за поведением пользователя.

Рассмотрим поведение пользователя в интернет-магазине в процессе выбора им товара определенной категории.

В данной ситуации, мы можем зафиксировать поведение потребителя, которое может проявиться в: 1) просмотре (можно зафиксировать время затраченное пользователем на анализ характеристик определенного товара); 2) отборе понравившихся товаров (помещение товаров в «корзину»); 3) покупке одной из марок товаров. В принципе, каждое из этих действий можно условно рассматривать как реализацию процесса выбора этим пользователем.

Полученную информацию невозможно перевести в числовую форму, так как у нас нет даже каких-либо оценок пользователей. Таким образом мы имеем количественные данные (характеристики товара $K^H(x_i)$, $i = \overline{1,n}$) на «входе» модели оценивания и качественную информацию (поведение пользователя) на «выходе». Поэтому применение классического метода параметрической идентификации математической модели (3) здесь невозможно и необходимо разработать принципиально новый подход к решению этой задачи. В качестве основы такого подхода предлагается использовать метод компараторной идентификации [4].

Рассмотрим способы реализации этого подхода. Пусть на основе интроспективного анализа множества объектов данной категории $X = \{x_1, x_2, ..., x_n\}$

пользователь реализовал одно из действий (поведений) описанных выше (например, выбрал товар $x_v \in X$). В соответствии с основными постулатами теории полезности и принятой гипотезе процесса выбора (2) можно записать, что

$$x_v \succ x_i \;\Leftrightarrow\; P(x_v) > P(x_i),\ \forall i = \overline{1,n},\ v \in I_n,\ i \neq v \quad (6)$$

или

$$P(x_v) - P(x_i) > 0,\ \forall i = \overline{1,n},\ v \in I_n,\ i \neq v. \quad (7)$$

Таким образом, получаем систему из $n-1$ линейных неравенств.

Следует заметить, что если на множестве альтернатив X установлены отношения полного или частичного порядка, то это обстоятельство легко формализуется путем добавления соответствующих ограничений в систему (7). Например, если предпочтения пользователя выглядят так: $x_1 \succ x_2 \sim x_3 \succ x_4$, то можно записать согласно (6), что $P(x_1) > P(x_2)$, $P(x_2) = P(x_3)$ и $P(x_3) > P(x_4)$.

Система неравенств (7) определяет выпуклый многогранник на гиперплоскости

$$\sum_{j=1}^{m} w_j = 1.$$

Любая точка $(w_1, w_2, ..., w_m)$, удовлетворяющая этой системе является решением задачи. Это означает, что задача идентификации параметров модели (2), (3) является некорректной по Тихонову (т. е. в общем случае, не имеет единственного решения). Для ее регуляризации примем в качестве единственного решения чебышевскую [5] или среднюю точки [4]. Нахождение решения этими методами сводится к решению задачи линейного программирования (ЛП) [4, 5], что не вызывает принципиальных трудностей. Единственное преобразование, которое необходимо сделать для этого — в системе линейных ограничений (7) заменить все знаки «>» на «≥» и добавить условия (4). Так как чебышевская точка является точкой, которая равноуклонена от граней области допустимых решений (ОДР) образуемой (4) и (7), т. е. решение «центрируется» относительно граней, а средняя точка представляет собой решение «центрированное» относительно вершин ОДР, то изменение знаков ограничений в (7) является чисто техническим приемом для сведения исходной задачи к задаче ЛП.

Экспериментальная проверка показала, что данные решения обладают высокой устойчивостью и имеют близкие к реальным значения [4].

Таким образом, в результате решения задачи параметрической идентификации модели выбора мы получим относительные весовые коэффициенты $(w_1, w_2, ..., w_m)$ важности для потребителя конкретных характеристик товаров из данной категории. Используя эти значения мы можем вычислить значения скалярных оценок $P(x_i)$, $i = \overline{1,n}$ для всех товаров этой категории по формуле (3), а затем ранжировать их на основе этих оценок в порядке убывания их «полезности» для потребителя, что даст возможность повысить релевантность выдачи рекомендаций.

3. Иллюстративный пример

Пусть в интернет-магазине имеется семь марок биноклей, т. е. $X = \{x_1, x_2, ..., x_7\}$ (табл. 1), из которых пользователь просмотрел только четыре. Наблюдая за поведением пользователя мы получили следующую информацию: отложил в «корзину» — x_3; просматривал характеристики x_2 (90 сек), x_5 (60 сек) и x_6 (30 сек). Предположим, что время просмотра характеристик биноклей определяет «интерес» пользователя к ним. Таким образом, его предпочтения могут быть представлены так:

$$x_3 \succ x_2 \succ x_5 \succ x_6. \quad (8)$$

На основе этой информации необходимо предсказать «ценность» (полезность) для пользователя марок биноклей x_1, x_4 и x_7, которые он не рассматривал. Эти предсказанные оценки послужат основанием выдачи рекомендаций пользователю в порядке убывания их «ценности» для него.

Пусть каждый бинокль описывается семью частными характеристиками (табл. 1.), которые определены на основании общих рекомендаций по их выбору. Эти характеристики

Таблица 1

Характеристики марок биноклей

x_i	*Марка бинокля*	k_1	k_2	k_3	k_4	k_5	k_6	k_7
x_1	Alpen Shasta Ridge 8.5x50	8.5	50	0.85	5700	5.90	4.0	98
x_2	Alpen Shasta Ridge 10x50	10.0	50	0.85	6100	5.00	4.0	94
x_3	Arsenal 12x50 Porro	12.0	50	0.95	4500	4.16	5.0	95
x_4	Barska X-Trail 30x80	30.0	80	2.28	5200	2.67	3.7	25
x_5	Bresser Spezial-Saturn 20x60	20.0	60	1.14	4500	3.00	9.0	52
x_6	Nikon Action EX 12x50 CF	12.0	50	1.00	7200	4.20	7.0	96
x_7	Nikon Prostaff 5 10x42	10.0	42	0.63	7000	4.20	5.0	98

Таблица 2

Нормированные характеристики биноклей

	$k_1^H(x_i)$	$k_2^H(x_i)$	$k_3^H(x_i)$	$k_4^H(x_i)$	$k_5^H(x_i)$	$k_6^H(x_i)$	$k_7^H(x_i)$	$P(x_i)$
w_i	0.102	0.102	0.102	0.256	0.102	0.141	0.195	---
x_1	0.00	0.21	0.87	0.56	1.00	0.94	1.00	0.683
x_2	0.07	0.21	0.87	0.41	0.72	0.94	0.95	0.613
x_3	0.16	0.21	0.81	1.00	0.46	0.75	0.96	0.716
x_4	1.00	1.00	0.00	0.74	0.00	1.00	0.00	0.534
x_5	0.53	0.47	0.69	1.00	0.10	0.00	0.37	0.511
x_6	0.16	0.21	0.78	0.00	0.47	0.38	0.97	0.408
x_7	0.07	0.00	1.00	0.07	0.47	0.75	1.00	0.476

следующие: — кратность увеличения определяет, насколько «ближе» к нам окажется наблюдаемый объект (в разах — чем больше, тем лучше); $k_4(x_i)$ — цена бинокля (в грн. — чем меньше, тем лучше); $k_5(x_i)$ — диаметр выходного зрачка определяет диаметр светового пучка, который попадает из бинокля в зрачок наблюдателя (в мм. — чем больше, тем лучше); $k_6(x_i)$ — минимальная дистанция фокусировки (в м. — чем меньше, тем лучше); $k_7(x_i)$ — ширина поля зрения на 1000 м. (в м. — чем больше, тем лучше).

Нормированные по формуле (5) характеристики биноклей представлены в табл. 2. В соответствии с информацией о предпочтениях потребителя (8) и принципом (6), а также с учетом формулы (3) и условий (4), система ограничений задачи может быть записана в следующим образом:

$$\begin{aligned} &P(x_3)-P(x_2)\ge 0,\\ &P(x_2)-P(x_5)\ge 0,\\ &P(x_5)-P(x_6)\ge 0,\\ &w_j\ge 0,\ j=\overline{1,7};\\ &\sum_{j=1}^{7} w_j = 1. \end{aligned} \qquad (9)$$

В результате вычисления чебышевской точки для системы линейных ограничений (9) получим значения весовых коэффициентов характеристик биноклей (строка w_i) и функций их полезности (столбец $P(x_i)$), которые представлены в табл. 2. Заметим, что все значения $P(x_i)$, $i=\overline{1,7}$ определяются по формуле (3). Таким образом, выдача рекомендаций пользователю осуществляется в порядке убывания полученных значений относительных оценок $P(x_i)$ и будет выглядеть так: x_1 ($P(x_1)$=0.683), x_4 ($P(x_4)$=0.534) и x_7 ($P(x_7)$=0.476) (см. табл. 2).

Выводы

Предложенный в работе подход может быть использован в качестве составной части при создании гибридной рекомендательной системы наряду с применением подхода основанного на контенте (content-based), основной идеей которого является выдача «близких» (например, с точки зрения расстояния Евклида) к выбранным пользователем объектов. Главным преимуществом подхода является применение неявных методов для получения информации о предпочтениях пользователя. Он позволяет частично решить проблему «холодного старта» присущей любому типу рекомендательных систем. Экспериментальные исследования подтвердили эффективность применения предложенного подхода.

Дальнейшие исследования должны быть направлены на разработку методов сегментации пользователей на основе полученных многофакторных оценок объектов и любой иной дополнительной информации о них.

Список литературы:

1. Рекомендательные системы / [Электронный ресурс]. — Режим доступа: http://vas3k.ru/blog/355/ **2.** *Jannach D., Zanker M., Felfernig A., Friedrich G.* Recommender Systems. An Introduction. — New York: Cambridge University Press, 2011. — 352 P. **3.** *Ricci F., Rokach L., Shapira B., Kantor P. B.* Recommender Systems Handbook. — Springer, 2011 — 845 p. **4.** *Петров К.Э., Крючковский В. В.* Компараторная структурно-параметрическая идентификация моделей скалярного многофакторного оценивания. — Херсон: Олди-плюс, 2009. — 294 с. **5.** *Зуховицкий С.И., Авдеева Л.И.* Линейное и выпуклое программирование. — М. : Наука, 1967. — 460 с.

Поступила в редколлегию 12.02.2018