ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 5(23), 2014, с. 55–65

УДК 004.9

Д. М. Понизовкин

Влияние меры сходства на результативность РС

Аннотация. В данной работе рассматриваются две основные задачи традиционных РС, вводятся понятия степень релевантности и релевантность между контентами пользователей или объектов, вычисление которых производится с помощью меры сходства. От выбора меры сходства зависит выполнение транзитивности отношения релевантности. В работе показано влияние транзитивности на значение критерия качества, который характеризует качество работы РС. Показано, что не всегда свойство транзитивности выполняется в традиционных РС.

Ключевые слова и фразы: рекомендательная система, мера сходства, транзитивность, релевантность контентов.

Введение

Традиционная рекомендательная система (далее РС) — это информационная система, целевые задачи которой состоят:

(1) в поиске N объектов таких, что степень релевантности их контентов и контента пользователя высока;
(2) в выявлении степени релевантности контента некоторого объекта РС контенту пользователя.

Для упрощения изложения будем считать, что между пользователями и их контентами существует взаимно-однозначное соответствие. Тоже верно и для объектов. Далее по тексту контент пользователя и пользователь или контент объекта и объект — синонимичные и взаимозаменяемые понятия. К примеру, говоря, что объекты релевантны, имеется в виду, что релевантны их контенты. Здесь и далее контент — это некоторая структура u, которая хранит информацию о пользователе, или t, которая хранит информацию об объекте. Множество контентов пользователей обозначим U, $|U| = |U|$, объектов —

T, $|T| = |T|$. Между любыми двумя контентами *выполняется отношение релевантности* $\mathbf{R}$, если *степень их релевантности высока*. Говоря о степени релевантности или релевантности пользователя или объекта, будем предполагать степень релевантности или релевантность их контентов. Степень релевантности u^i и t^j задается самим пользователем РС оценкой v^i_j по некоторой шкале $\mathcal{S}$ в момент работы с системой, однако не для каждой пары (u^i, t^j) $\exists$ v^i_j. Степень релевантности может быть вычислена алгоритмически через задание *меры сходства* sim — функции, сопоставляющей паре контентов вещественное число, определяющее степень релевантности в области определения $\mathcal{D}_{\text{sim}}$ меры сходства. Система (V, e, sim) ранжирует множество V: $\text{sim}(v_i, e) \geq \text{sim}(v_{i+1}, e)$.

Первая целевая задача РС называется top-N [1]. Она заключается в формировании подмножества $T^a_{\text{top}} = \{t \in T' \subset T | t\mathbf{R}u^a\}$, $|T^a_{\text{top}}| = N$, u^a — активный пользователь системы, для которого в данный момент решается задача РС.

Вторая целевая задача РС называется задачей *прогнозирования неизвестной оценки* v^a_p [2] и определяется как вычисление такой оценки vp^a_p , что $vp^a_p \approx v^a_p$.

Для оценки качества работы РС проводится тестирование, после которого производится расчет *критерия оценки*. Критерий оценки — функция, сопоставляющая результату вещественное число, по которому можно судить о качестве РС. Множество входных данных, как правило, разбивается на два подмножества: обучающее (будем индексировать его и его элементы символом 0), на котором производится тест, и тестовое (будем индексировать его и его элементы символом $\times$), с котором производится сравнение полученного результата.

1. Задача top-N в традиционных РС

1.1. Модель

Неформально модель задачи top-N традиционных РС определяется следующим *эвристическим утверждением*: если пользователю «нравится» объект t^1 и объект t^2 «похож» на объект t^1, то пользователю «понравится» объект t^2 [1]. Понятие «похож» определяется традиционными РС на основании значений мер сходства $\mathcal{D}_{\text{sim}} = T \times T$, понятие «нравится» не определено, так как не введена $\mathcal{D}_{\text{sim}} = U \times T$. Следуя введенной в статье терминологии, утверждение примет вид:

$$u^a \mathbf{R} t^i \wedge t^i \mathbf{R} t^j \Rightarrow u^a \mathbf{R} t^j \tag{1}$$

Формула (1) описывает транзитивность отношения релевантности.

Необходимые данные для решения задачи и расчета критерия оценки:

(1) $T_0^a = \{t_0^j | t_0^j \mathbf{R} u^a\}$
(2) $T_\times^a = \{t_\times^j | t_\times^j \mathbf{R} u^a\}$
(3) $T_0^a \bigcap T_\times^a = \varnothing$

1.2. Схема решения

Неформальное описание схемы решения определяется следующим образом: найти N объектов, наиболее «схожих» с теми, которые нравятся пользователю.

Схема решения:

$$T_{\text{top}}^a = \{t_1, ..., t_N\} \subset (T \setminus T_0^a, T_0^a, \sum_{t_0 \in T_0^a} \text{sim}) \tag{2}$$

Модель и решение реализованы, к примеру, компанией Amazon [3].

1.3. Пример традиционного решения задачи top-N

Контентом пользователя является вектор, отображающий информацию о том, какие объекты пользователь предпочел (поставил высокую оценку) за время работы с РС. $u^a = (v_1^a, ..., v_{|T|}^a)$, где:

$$v_j^a = \begin{cases} 1, & \text{если } u^i \mathbf{R} t^j \\ 0, & \text{иначе.} \end{cases}$$

$t^j \in T_0$. К примеру, для системы Amazon, если $v_j^a = 1$, то пользователь приобрел товар j.

Контент объекта $t^j = (ct_1^j, .., ct_{nt}^j)$ — вектор некоторых характеристик. Множество характеристик объектов и их значения зависят от системы, и для описания решения их определение не является обязательным.

Типичной мерой сходства, используемой при решении задачи top-N является $\text{sim}(t^i, t^j) = \cos(\angle(t^i, t^j))$ — косинус угла между векторами [3].

Модель РС задается матрицей $\mathcal{M}$ мер сходства размерности $|T| \times |T|$:

$$\mathcal{M}_{ij} = \begin{cases} \mathrm{sim}(t^i, t^j), & i \neq j \\ 0, & \text{иначе} \end{cases}$$

Решение основано на (2) и принимает вид:

$$T^a_{\mathrm{top}} = \left\{ t \middle| \sum_{t_0} \mathrm{sim}(t, t_0) \to \max \right\}$$

Шаги решения:

(1) $s = \mathcal{M} \times u^a = (s_1, ..., s_{|T|})$, где

$$s_j = \begin{cases} s_j, & \text{если } s_j \in (s, 0, s_i) \\ 0, & \text{иначе} \end{cases}$$

в вектор s входит N максимальных элементов, остальные заменяются на ноль

(2) $T^a_{\mathrm{top}} = \{t^j | s_j \neq 0\}$.

Algorithm 1. Построение матрицы мер сходства модели

```
for i = 1, i < |T| do                          ▷ Построение модели
    for j = i, j < |T| do
        M_ij = sim(t^i, t^j)
        M_ji = M_ij
        j = j + 1
    end for
    i = i + 1
end for
```

Algorithm 2. Решение задачи *top-N*

```
s = (s_1, ..., s_|T|) = M × u^a
T^a_top = {t^j | s_j ∈ (s, 0, s_j)}
```

1.4. Схема оценки качества решения задачи top-N. Примеры

Для того, чтобы определить качество решения задачи top-N, необходимо определить число объектов результирующей выборки, релевантных пользователю: $|\{t^j_{\text{top}} \in T^a_{\text{top}} | u^a \mathbf{R} t^j_{\text{top}}\}|$. Описанная модель и схема решения не использует $\mathcal{D}_{\text{sim}} = U \times T$, поэтому определение качества основано на (1): $t^i_\times \mathbf{R} t^j_{\text{top}} \wedge u^a \mathbf{R} t^i_\times \Rightarrow u^a \mathbf{R} t^j_{\text{top}}$. Существующие критерии E_{top} прямо-пропорциональны числу $|\{t_{\text{top}} | t_{\text{top}} \mathbf{R} t_\times\}|$.

Примеры критериев E_{top}:

- прежде, чем задать примеры, введем функцию

$$r(i) = \begin{cases} 1, & \exists\, t_\times \; t^i_{\text{top}} \mathbf{R} t_\times \\ 0, & \text{иначе.} \end{cases}$$

- Точность $P = \frac{1}{N} \cdot \sum\limits_{i=1}^{N} r(i)$;
- $P@K = \frac{1}{K} \cdot \sum\limits_{i=1}^{K} r(i)$;
- $AveP = \dfrac{1}{\sum\limits_{l} r(i)} \sum\limits_{K=1}^{N} P@K \cdot r(K)$;

Результат решения качественный, когда $E_{\text{top}} \to \max$. При нарушении свойства транзитивности релевантности $E_{\text{top}} \not\to \max$.

1.5. Надежность решения и транзитивность релевантности

Схема решения *не гарантирует* $E_{\text{top}} \to \max$, так как не гарантирует выполнения (1) в общем случае:

- пользователь предпочитает различные объекты;
- с течением времени вкусы пользователя могут смениться.

Эти причины могут привести к тому, что $T_0 \neq \{t_0 | t_0 \mathbf{R} t_\times\}$, за счет чего $E_{\text{top}} \not\to \max$.

2. Задача прогнозирования традиционных РС

2.1. Модель

Неформально модель задачи прогнозирования в традиционных РС определяется следующим *эвристическим утверждением*: пользователи с «похожими» предпочтениями в прошлом, будут иметь

«похожие» предпочтения в будущем. Предпочтения для задачи прогнозирования выражаются оценкой, а понятие «похож» определяется традиционными РС на основании значений мер сходства $\mathcal{D}_{\text{sim}} = U \times U$. Следуя введенной в статье терминологии, утверждение примет вид:

$$u^1 \mathbf{R} u^2 \Leftrightarrow \forall t^i : v_i^1 \approx v_i^2 \tag{3}$$

Модель заключается в выявлении множества *соседей* активного пользователя $\mathcal{N}_p^a$:

$$\mathcal{N}_p^a = \{u^i | u^i \mathbf{R} u^a\} \wedge \exists v_p^i, i = 1..N$$

На основании оценок этих пользователей составляется прогнозная оценка.

Это направление разрабатывалось, к примеру, разрабатывались исследователями, принимавшими участие в соревновании Netflix Prize. Необходимые данные для решения задачи и расчета критерия оценки:

(1) $U_0^a = \{v_j^a\}$
(2) $U_\times^a = \{v_j^a\}$
(3) $U_0^a \bigcap U_\times^a = \varnothing$

2.2. Схема решения

$$\mathcal{N}_p^a = \{u^1, ..., u^N\} \subset (U, u^a, \text{sim}), \exists v_p^i, i = 1..N$$

Прогнозная оценка:

$$vp_p^a \approx pr\{v_p^i\}, u^i \in \mathcal{N}_p^a$$

pr — некоторая функция, сопоставляющая множеству оценок прогнозную оценку. Примеры функции pr:

- Средняя оценка:

$$\frac{1}{\mathcal{N}_p^a} \cdot \sum_{u^i \in \mathcal{N}_p^a} v_p^i$$

.

- Средняя взвешенная оценка:

$$\frac{\sum\limits_{i \in \mathcal{N}_p^a} \mathrm{sim}(u^a, u^i) \cdot v_p^i}{\sum\limits_{u^i \in \mathcal{N}_p^a} |\mathrm{sim}(u^a, u^i)|}$$

2.3. Пример традиционного решения задачи прогнозирования

Algorithm 3. Построение множества соседей

$\mathcal{N}_p^a = \varnothing$
for $i = 1, |U|$ **do**
  if $u^a \mathbf{R} u^i \wedge \exists\, v_p^i$ **then**
    $\mathcal{N}_p^a = \mathcal{N}_p^a \bigcup \{u^i\}$
    $i = i + 1$
  end if
  if $i \geq N$ **then**
    выход из цикла
  end if
end for

Algorithm 4. Решение задачи прогнозирования при f, являющей простой взвешенной суммой

$vp_p^a = 0$
$divider = 0$
for $i = 1, N$ **do**
  $s = |\mathrm{sim}(u^a, u^i)|$
  $vp_p^a = vp_p^a + s \cdot v_p^u$
  $divider = divider + s$
  $i = i + 1$
end for
$vp_p^a = \frac{vp_p^a}{divider}$

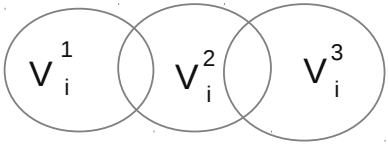

Рис. 1. Транзитивность «близких» оценок

2.4. Схема оценки качества, критерий E_p

Расчет оценки качества основан на (3): так как прогнозная оценка строится на оценках релевантных пользователей, то она должна быть приблизительно равна реальной. Для оценки качества определяется разность прогнозной и реальной. Примеры критерия E_p:

(1) $MAE = \frac{1}{M} \sum\limits_{p=1}^{M} |vp_p^a - v_p^a|$

(2) $RMSE = \sqrt{\frac{1}{M} \sum\limits_{p=1}^{M} (vp_p^a - v_p^a)^2}$

где M — число прогнозных оценок.

Результат решения является качественным, если $E_p \to \min$. Традиционные РС основывают решение на предположении, что, если $v_p^a \approx v_p^i \approx v_p^j \approx pr(\{v_p^a, v_p^j\}) \Rightarrow |vp_p^a - v_p^a| \to \min$. Предполагается, что модель и схема решения гарантирует выполнение $E_p \to \min$, так как

$$u^a \mathbf{R} u^i \wedge u^a \mathbf{R} u^j \Rightarrow u^i \mathbf{R} u^j \tag{4}$$

2.5. Надежность решения и транзитивность релевантности

Схема решения, в общем случае, не гарантирует выполнение $E_p \to \min$:

- Отношения $\approx$ не транзитивно: $\forall t^i : v_i^1 \approx v_i^2 \wedge \forall t^i : v_i^2 \approx v_i^3 \not\Rightarrow \forall t^i : v_i^2 \approx v_i^3$ Эта ситуация проиллюстрирована на Рис 1. Таким образом, если $u^1, u^2 \in \mathcal{N}^a \not\Rightarrow \forall t^i : v_i^1 \approx v_i^2 \Rightarrow E_P \not\to \min$
- Не всякая мера сходства гарантирует выполнение транзитивности. К примеру, в качестве функции sim часто используются коэффициенты корреляции, для которых транзитивность может не выполняться [4, 5]:

$$u^1 \mathbf{R} u^2 \wedge u^2 \mathbf{R} u^3 \not\Rightarrow u^2 \mathbf{R} u^3, u^i \in \mathcal{N}$$

Таблица 1. Сравнение качества традиционных РС

Тип РС	MAE	NMAE	RMSE
Традиционная РС, Пирсон	0.85	0.21	1.09
Реформированная, Хэмминг	0.65	0.16	0.96

- Если предпочтения пользователей схожи, то это не означает, что их оценки схожи [7]. Данное обстоятельство является следствием того, что характеры пользователей могут отличаться. Лояльные пользователи ставят оценки не ниже определенной, критики — наоборот, поэтому разница оценок может быть существенной, несмотря на релевантность пользователей.

3. Сравнение качества традиционных РС при использовании различных мер сходства

Для проведения тестов РС была взята база данных MovieLens 1M. Описание данных:

- число пользователей — 6040;
- число объектов — 3952;
- $v_j^i \in \{1, 2, 3, 4, 5\}$;

На этих данных была решена задача прогнозирования в традиционной модели РС с использованием коэффициента корреляции Пирсона в качестве меры сходства. Полученные результаты сравнивались с традиционной РС, работающей с контентами, которые представляют собой нечеткие множества. Описание подобного представления данных приведено в [6].Для этого была проведена небольшое реформирование контентов пользователей: $v_j^i = \frac{(v_j^i - 1)}{4}$ и использованием расстояния Хэмминга в качестве меры сходства. *Кроме данной реформации данных никаких других настроек системы с целью улучшения результата не производилось.*

4. Заключение

Для получения хороших показателей критериев качества оценки РС, необходимо, чтобы выполнялось свойство транзитивности релевантности контентов. Однако данное свойство выполняется не всегда в традиционных РС. Качество традиционных РС можно повысить

с помощью небольшой реформации данных и использовании мер сходства, обладающих метрическими свойствами.

Список литературы

[1] M. Deshpande, G. Karypis, “Item-based top-N recommendation algorithms”, *ACM Transactions on Information System*, **22**:1, pp. 143–177 ↑ 56.

[2] X. Su, T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques”, *Advances in Artificial Intelligence*, **2009**, pp. 19 ↑ 56.

[3] G. Linden, B. Smith, J. York, “Amazon.com recommendations: item-to-item collaborative filtering”, *IEEE Internet Computing*, **7**:1 (2003), pp. 76–80 ↑ 57.

[4] A. E. C. Sotos, S. Vanhoof, W. V. Noortgate, P. Oughena, “The Non-Transitivity of Pearson’s Correlation Coefficient: An Educational Perspective”, Proceedings of the 56th Session of the ISI, *Bulletin of the ISI*, **62**, pp. 4609–4613 ↑ 62.

[5] P. Burger, *Non-Transitivity Property of Correlation*, http:// ec2-54-245-224-94.us-west-2.compute.amazonaws.com/wordpress/2013/05/24/non-transitivity-property-of-correlation/ ↑ 62.

[6] С. А. Амелькин, Д. П. Понизовкин, «Математическая модель задачи top-N для контентных рекомендательных систем», *Известия МГТУ МАМИ*, **2**, с. 26–31 ↑ 63.

[7] R. Jin, L. Si, C. Zhai, J. Callan, “Collaborative filtering with decoupled models for preferences and ratings”, Proc. of the 12th international conference on Information and knowledge management, pp. 309–316 ↑ 63.

Рекомендовал к публикации *д.т.н. В. И. Гурман*

Об авторе:

Денис Михайлович Понизовкин

Автор статьи — аспирант, интересом исследований которого является область рекомендательных систем.

e-mail: denis.ponizovkin@gmail.com

Образец ссылки на эту публикацию:

Д. М. Понизовкин. *Влияние меры сходства на результативность РС* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. 5, № 5(23), с. 55–65.

URL http://psta.psiras.ru/read/psta2014_5_55-65.pdf

Denis Ponyzovkin. *Quality of recommender systems and transitivity of content's relevance.*

Abstract. In this paper, we propose relevance and degree of relevance of user's or item's contents. Heuristic motivations of traditional recommender systems assume that there is transitivity property of relevance. We introduce importance of transitivity property, its influence on values of evaluation metrics and show that transitivity not performed in general case for traditional recommender systems. (*in Russian*).

Key Words and Phrases: traditional recommender system, similarity measure, transitivity, content's relevance, evaluation metric.