DOI 10.37882/2223-2966.2023.07.07

АЛГОРИТМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЯ ДЛЯ ПРОГНОЗА РЕЙТИНГА ПРОДУКТОВ ПРИ ОТСУТСТВУЮЩИХ ОЦЕНКАХ НА ОСНОВЕ МЕРЫ СХОДСТВА СТАТИСТИЧЕСКОЙ ИМПЛИКАЦИИ

DECISION SUPPORT ALGORITHM FOR PREDICTION OF THE RATING OF PRODUCTS WITH NO RATINGS BASED ON THE MEASURE OF SIMILARITY OF THE STATISTICAL IMPLICATION

Vo Thi Huyen Trang
I. Kvyatkovskaya
Tran Quoc Toan

Summary: The measure of similarity plays an important role in the recommender system of collaborative filtering based on users, since it directly affects the results of recommender systems. To determine the measure of similarity between two users in a recommender system, many solutions are proposed, such as: using statistical correlation, using cosine distance between two vectors, using association rules, etc… In this paper, he proposes a measure of similarity based on the statistical implication analysis method. The measure of similarity between two users is determined based on the sum of the statistical implication distances of the association rules, which tend to favor counter-examples (the number of objects that satisfy property *a* but not property *b* in the association rule $a \rightarrow b$) generated from the ranking data of the two users.

Keywords: similarity measure; recommender system; statistical implication analysis; statistical implication intensity measure; product rating; rating matrix; association rule.

Во Тхи Хуен Чанг
Аспирант, ФГБОУ ВО «Астраханский государственный технический университет»
vthtrang@mail.ru
Квятковская Ирина Юрьевна
Доктор технических наук, профессор, ФГБОУ ВО «Астраханский государственный технический университет»
i.kvyatkovskaya@astu.org
Чан Куок Тоан
Кандидат технических наук, ФГБОУ ВО «Астраханский государственный технический университет»
hoaivan219@mail.ru

Аннотация. Мера сходства играет важную роль в рекомендательной системе совместной фильтрации, основанной на прогнозировании поведения пользователя, поскольку она напрямую влияет на фактор эффекта рекомендаций такой системы. Определить меру сходства между двумя пользователями в рамках рекомендательной системы можно на основе различных алгоритмов, таких, к примеру, как: расчет статистической корреляции, использование косинусного расстояния между двумя векторами, применение ассоциативных правил и т.д. В данной статье авторы рассматривают оценку меры сходства, основанную на методе анализа статистической импликации. Мера сходства между предпочтениями двух пользователей вычисляется с учетом суммы расстояний статистической импликации ассоциативных правил, которые характеризуют противоположные предпочтения — т.н. контрпримеры (число объектов, которые удовлетворяют свойству *a*, но не удовлетворяют свойству *b*, в контексте ассоциативного правила $a \rightarrow b$), созданным на основе данных ранжирования предпочтений двух пользователей.

Ключевые слова: мера сходства; рекомендательная система; анализ статистической импликации; мера интенсивности статистической импликации; рейтинг продуктов; рейтинговая матрица; ассоциативное правило.

Мера сходства, определяемая на базе интенсивности статистической импликации, — это мера, используемая для определения значения сходства между двумя пользователями на основе некоторого набора ассоциативных правил, построенных посредством сравнения рейтинга двух пользователей, и показателя интенсивности статистической импликации. Мера сходства интенсивности статистической импликации между двумя пользователями (u_a и u_b) определяется по следующей формуле:

$$SIS(u_a, u_b) = 1 - \sum_{i=1}^{k} II(r_i) \,/\, k, \qquad (1)$$

где $SIS(u_a, u_b)$ — мера сходства между двумя пользователями u_a, u_b; $I(r_i)$ — величина интенсивности статистической импликации ассоциативного правила r_i; k — количество правил ассоциации в наборе правил ассоциации, созданном на основе данных ранжирования двух пользователей u_a, u_b.

Мера интенсивности статистической импликации $\varphi(a,b)$ правила $a \rightarrow b$ определяется по формуле [1]:

$$\varphi(a,b) = 1 - \sum_{k=\max(0, n_A - n_B)}^{n_{A\bar{B}}} \frac{C_{n_B}^{n_A - k} C_{n-n_B}^{k}}{C_n^{n_A}}, \qquad (2)$$

где n — количество пользователей, $n_A = |A|$ и $n_B = |B|$ — количество элементов множеств A и B; $n_{A\bar{B}} = |A \cap \bar{B}|$ — число контрпримеров (это число объектов, которые удовлетворяют свойству a, но не удовлетворяют свойству b).

Псевдокод алгоритма для определения меры сходства статистической импликации между двумя пользователями u_a, u_b выглядит следующим образом:

Входные данные — это рейтинговые данные для продуктов двух пользователей u_a, u_b.

Выходные данные: значение сходства между двумя пользователями u_a, u_b.

Последовательность действий:

Шаг 1. Необходимо сгенерировать ассоциативные правила на основе рейтинговой матрицы пользователя.

Шаг 2. Следует выбрать ассоциативные правила для двух пользователей — u_a, u_b:
- — выбрать продукты, оцененные пользователями u_a: $\overline{I_{ua}}$;
- — выбрать продукты, не оцененные пользователями u_b: $\overline{I_{ub}}$;
- — выбрать ассоциативные правила вида $X \rightarrow Y$, где $X \in I_{ua}; Y \in \overline{I_{ub}}$ и $X \cap Y = \varnothing$.

Шаг 3. Вычисляются значения параметров $n, n_A, n_B, n_{A\bar{B}}$ для каждого правила из выбранного набора ассоциативных правил.

Шаг 4. Определяется величина интенсивности статистической импликации для выбранного набора ассоциативных правил.

Шаг 5. Определяется степень сходства между двумя пользователями — u_a, u_b, в частности:
- — вычисляется среднее значение интенсивности статистической импликации для набора правил ассоциации $\bar{S}$;
- — определяется значение сходства между двумя пользователями u_a, u_b: $SIS(u_a, u_b) = 1 - \bar{S}$.

Примечание: $0 \le II(r_i) \le 1$, то $0 \le i = \sum_{i=1}^{k} II(r_i) / k \le 1$.

Таким образом, $SIS(u_a, u_b) \in [0,1]$.

Пример 1. Рассмотрим рейтинговую матрицу двух пользователей $\{u_1, u_2\}$, оценивающих четыре продукта $\{i_1, i_2, i_3, i_4\}$ (см. табл. 1).

Таблица 1.

Рейтинговая матрица двух пользователей

	i_1	i_2	i_3	i_4
u_1	0	4	4	1
u_2	0	0	4	0

Источник: составлено авторами на основе проведенного исследования.

Первым шагом является создание ассоциативных правил из матрицы ранжирования. Затем производится выбор правила для двух пользователей — u_1, u_2. Ассоциативные правила двух пользователей сведены в таблицу 2.

Таблица 2.

Ассоциативные правила двух пользователей u_1, u_2

№	Ассоциативные правила
1	$\{i_2=4\} \rightarrow \{i_1=0\}$
2	$\{i_4=1\} \rightarrow \{i_1=0\}$
3	$\{i_3=4\} \rightarrow \{i_1=0\}$
4	$\{i_3=4\} \rightarrow \{i_2=0\}$
5	$\{i_3=4\} \rightarrow \{i_4=0\}$
6	$\{i_2=4, i_4=1\} \rightarrow \{i_1=0\}$
7	$\{i_2=4, i_3=4\} \rightarrow \{i_1=0\}$
8	$\{i_3=4, i_4=1\} \rightarrow \{i_1=0\}$
9	$\{i_2=4, i_3=4, i_4=1\} \rightarrow \{i_1=0\}$

Источник: составлено авторами на основе проведенного исследования.

Следующим шагом является определение параметров $n, n_A, n_B, n_{A\bar{B}}$ для каждого правила ассоциации и вычисление величины интенсивности статистической импликации на основе этих параметров. Результаты этого шага представлены в таблице 3.

Таблица 3.

Значение параметров и величина интенсивности статистической импликации каждого правила ассоциации

№	Правило ассоциации	n	n_A	n_B	$n_{A\bar{B}}$	Интенсивность
1	$\{i_2=4\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49
2	$\{i_4=1\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49
3	$\{i_3=4\} \rightarrow \{i_1=0\}$	2	2	1	1	0,38
4	$\{i_3=4\} \rightarrow \{i_2=0\}$	2	2	1	1	0,38
5	$\{i_3=4\} \rightarrow \{i_4=0\}$	2	2	1	1	0,38
6	$\{i_2=4, i_4=1\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49
7	$\{i_2=4, i_3=4\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49
8	$\{i_3=4, i_4=1\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49
9	$\{i_2=4, i_3=4, i_4=1\} \rightarrow \{i_1=0\}$	2	1	1	0	0,49

Источник: составлено авторами на основе проведенного исследования.

Наконец, сходство между пользователями u_1, u_2 определяется следующим образом: $SIS(u_a, u_b) = 1 - 0{,}45 = 0{,}55$.

Алгоритм поддержки принятия решения для прогноза рейтинга продуктов при отсутствующих оценках на основе меры сходства статистической импликации

Рекомендательная модель совместной фильтрации, основанная на мере сходства статистической импликации, описывается следующим образом:

$$CFRS = <U, I, R, F>, \quad (3)$$

где $U = \{u_1, u_2, \ldots, u_n\}$ — множество n пользователей системы; $I = \{i_1, i_2, \ldots, i_m\}$ — множество m продуктов системы; $R = \{r_{j,k}\}$ — рейтинговая матрица пользователей для продуктов, где каждая строка представляет пользователя $u_j (1 \le j \le n)$, каждый столбец представляет продукт $i_k (1 \le k \le m)$, $r_{j,k}$ — значение рейтинга пользователя u_j для продукта i_k; $F : U \times I \times R \to I_{u_a}$ — вычислительная функция, позволяющая определить продукты, которые необходимо рекомендовать пользователю $u_a \in U$: $I_{u_a} = \{i^1_{u_a}, \ldots, i^N_{u_a}\}$.

Псевдокод алгоритма поддержки принятия решения для прогноза рейтинга продуктов при отсутствующих оценках на основе меры сходства статистической импликации выглядит следующим образом:

Входные данные: набор пользователей U, набор продуктов I, рейтинговая матрица пользователей для продуктов R, при этом нужная пользователю рекомендация обозначена как u_a.

Выход: продукты, которые необходимо рекомендовать пользователю $u_a \in U$: $I_{u_a} = \{i^1_{u_a}, \ldots, i^N_{u_a}\}$.

Последовательность действий:

Шаг 1. Необходимо определить список k пользователей, похожих на пользователя u_a.

Для каждого пользователя $u_i \in U$ следует:
- определить значение меры сходства между u_a и u_i на основе меры сходства статистической импликации: $SIS(u_a, u_i)$;
- выстроить список пользователей по убыванию значения сходства;
- выбрать первых k пользователей с наибольшим значением меры сходства ($N(u_a)$).

Шаг 2. Рассчитываются прогнозируемые рейтинги продуктов:
- определяются продукты, которые пользователь u_a не оценил ($r_{a,k} = \varnothing$);
- рассчитываются прогнозируемые рейтинги для этих продуктов по формуле:

$$\overline{r_{a,k}} = \frac{1}{\sum_{i \in N(a)} s_{a,i}} \sum_{i \in N(a)} s_{a,i} r_{i,k},$$

где $s_{a,i}$ — значения сходства между u_a и u_i; $r_{i,k}$ — значение рейтинга пользователя u_i для продукта i_k.

Шаг 3. Выбрать продукты для рекомендации пользователю u_a:
- отсортировать продукты по убыванию значения их прогнозируемых рейтингов ;
- выбрать продукты с наивысшим значением прогнозируемых рейтингов, для того чтобы представить их пользователю u_a.

Пример 2. Чтобы более наглядно проследить последовательность действий в рамках алгоритма, предположим, что система предлагает пользователям выбрать восемь условных продуктов (от i_1 до i_8), и в настоящее время в системе есть 10 пользователей (от u_1 до u_{10}), которым присвоены рейтинги продуктов. Товары оцениваются по шкале от 1 до 5 (где «1» — самая низкая оценка; «5» — самая высокая оценка; а «?» — продукт, не имеющий рейтинга от пользователя). Система должна рекомендовать продукты новому пользователю u_a с коэффициентом k, равным 4 (то есть расчет производится для четырех похожих пользователей), как показано в таблице 4.

Таблица 4.

Рейтинговая матрица между пользователями и продуктами

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4,0	4,0	1,0	2,0	2,0	?	?
u_2	3,0	?	?	?	5,0	1,0	?	?
u_3	4,0	?	?	3,0	2,0	2,0	?	3,0
u_4	3,0	?	3,0	2,0	1,0	?	5,0	4,0
u_5	1,0	1,0	?	?	?	?	?	1,0
u_6	?	1,0	?	?	?	1,0	?	1,0
u_7	1,0	4,0	?	2,0	?	4,0	4,0	?
u_8	5,0	?	4,0	3,0	?	2,0	3,0	1,0
u_9	?	1,0	3,0	?	?	?	1,0	?
u_{10}	2,0	?	?	?	?	3,0	2,0	1,0
u_a	?	?	4,0	2,0	?	1,0	?	3,0

Источник: составлено авторами на основе проведенного исследования.

Исходя из описанного выше требования рекомендательная система определяет список пользователей, похожих на u_a, опираясь на меру сходства статистической

импликации, включая: u_1, u_4, u_7, u_8, как показано на рисунке 1:

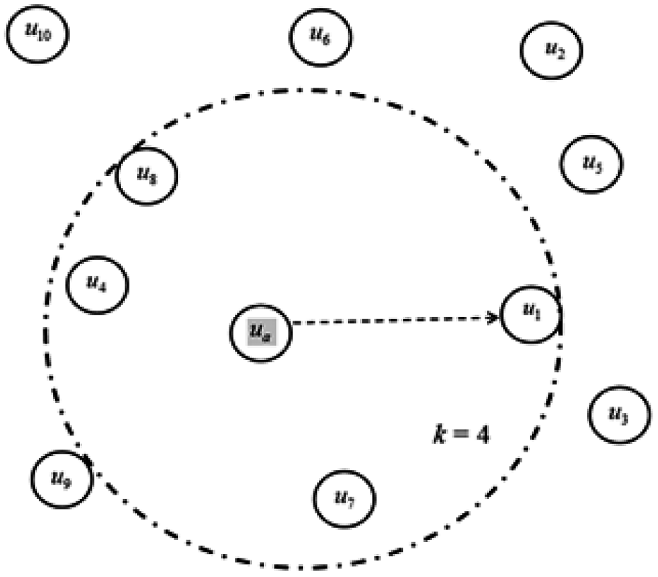

Рис. 1. Определение списка пользователей, похожих на пользователя u_a при $k = 4$

Источник: составлено авторами на основе проведенного исследования.

В таблице 5 приведен расчет списка прогнозируемых продуктов для пользователя u_a.

Таблица 5.

Расчет списка прогнозируемых продуктов для пользователя u_a

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_a	?	?	4,0	2,0	?	1,0	?	3,0
r_a	3,0	4,0			1,5		4,0	

Источник: составлено авторами на основе проведенного исследования.

На основе перечня похожих пользователей составляется система рекомендаций и осуществляется вычисление значения рейтинга для продуктов, рекомендуемых пользователю u_a и продуктов без рейтинга, и направляет пользователю u_a рекомендуемые продукты (включая: i_1, i_2, i_7), как показано в табл. 5.

Итак, в данной статье авторами была рассмотрена мера сходства для рекомендательной системы на основе метода анализа статистической импликации. Кроме того, разработан алгоритм поддержки принятия решения для прогноза рейтинга продуктов при отсутствующих оценках с учетом меры сходства статистической импликации.

ЛИТЕРАТУРА

1. Квятковская И.Ю. Модель и алгоритм поддержки принятия решения по выбору продуктов для рекомендации пользователю на основе метода анализа статистической импликации / И.Ю. Квятковская, Во Тхи Хуен Чанг, Чан Куок Тоан // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2023. — № 2. — С. 116–124. — URL: https://doi.org/10.24143/2072-9502-2023-2-116-124. EDN UHNZRL. — Текст электронный.
2. David H. Glass. Confirmation measures of association rule interestingness // Knowledge-Based Systems. — 2013. — № 44. — Pp. 65–77.
3. Шуршев В.Ф. Модель системы поддержки принятия решений на основе рассуждений по прецедентам / В.Ф. Шуршев, Г.А. Кочкин, В.Р. Кочкина // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2013. — № 2. — С. 175–183.
4. Kvyatkovskaya I.Y. Methodology of a support of making management decisions for poorly structured problems / I.Y. Kvyatkovskaya, V.F. Shurshev, M.B. Frenkel // Communications in Computer and Information Science. — 2015. — Т. 535. — С. 278–291.
5. Gras R. An overview of the Statistical Implicative Analysis (SIA) development / R. Gras, P. Kuntz // Statistical Implicative Analysis — Studies in Computational Intelligence, Springer-Verlag. — 2008. — № 127. — Pp. 11–40.
6. Gras R. Notion of Implicative Fields in Statistical Implicative Analysis / R. Gras, P. Kuntz, N. Greffard // VIII Colloque International — VIII International Conference A.S.I. Analyse Statistique Implicative — Statistical Implicative Analysis Radès (Tunisie). — Novembre. — 2015. — Pp. 29–46.

Журнал «Современная наука: актуальные проблемы теории и практики»