УДК 004.78

АЛГОРИТМ СИНГУЛЯРНОГО РАЗЛОЖЕНИЯ МАТРИЦ В МЕТОДЕ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ПРИ ПОСТРОЕНИИ СИСТЕМЫ РЕКОМЕНДАЦИЙ И ПРОГНОЗИРОВАНИЯ

Феоктистов Д.Г., Романовский Н.Д.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Ролич О.Ч. – канд. техн. наук, доцент, доцент кафедры ПИКС

Аннотация. Рассматривается алгоритм сингулярного разложения матриц в качестве одного из основных методов оптимизации метода коллаборативной фильтрации и анализ его воздействия на входные данные.

Ключевые слова: системы рекомендаций, разложение матриц, коллаборативная фильтрация

Введение. В информационных системах актуальна задача предоставления рекомендаций пользователям на основе имеющихся данных о совершенных действиях и результатах обратной связи. Решением данной задачи является построение рекомендательных систем. Одним из основных подходов к построению такого рода систем является использование метода коллаборативной фильтрации. В частности, оптимизация данного метода достигается использованием сингулярного разложения матриц.

Основная часть. Метод коллаборативной фильтрации основан на построении матрицы (графа), отражающей модель взаимодействий пользователей и товаров. Рекомендации (прогноз) для целевого пользователя формируются на основании вычисления некой меры похожести по всем накопленным данным [1].

Генеральная совокупность состоит из набора пользователей, который будет обозначен как U и набора условных товаров I.

Тогда U_i – i-й элемент набора пользователей, где $i = 1,2,\ldots n$;

I_j – j-й элемент набора товаров, где $j = 1,2,\ldots m$;

$r_{i,j}$ – оценка взаимодействия i-го пользователя с j-м товаром;

Данную матрицу можно в общем случае представить в виде таблицы *1*.

Таблица *1* – Таблица отношений

	I_1	I_2	…	I_m
U_1	$r_{1,1}$	$r_{1,2}$	…	$r_{1,m}$
U_2	$r_{2,1}$	$r_{2,2}$	…	$r_{2,m}$
…	…	…	…	…
U_n	$r_{n,1}$	$r_{n,2}$	…	$r_{n,m}$

Для вычисления меры схожести оценок пользователей воспользуемся косинусной мерой [2], которая в общем случае вычисляется по формуле *1*:

$$\boldsymbol{cos(\overline{x},\overline{y}) = \frac{\overline{x}\cdot\overline{y}}{||\overline{x}||\cdot||\overline{y}||}}. \quad (1)$$

Используя данную формулу, вычислим меру схожести $m_{k,l}$ оценок пользователя относительно остальных по формуле *2*. Пусть U_k – вектор оценок рассматриваемого пользователя ($U_k \in U, k = 1, 2, \ldots, n$). Тогда U_l – вектор оценок пользователя множества U, для которого выполняются условия $l \neq k, l = 1, 2, \ldots, n$.

$$\boldsymbol{m_{k,l} = co\,s(U_{k,}U_l)}. \quad (2)$$

Составляя вектор M_k, состоящий из элементов $m_{k,l}$, где $k = const$, а l изменяется от 1 до n, причем $l \neq k$, выделим вектор $L_k \in M_k$ пользователей, меры схожести которых наибольшим образом коррелируют с элементами U_k.

Затем для каждого из пользователей умножим его оценки на вычисленную величину меры $m_{k,l}$. Таким образом, оценки более «похожих» пользователей будут сильнее влиять на итоговую позицию продукта и из продуктов посчитаем сумму калиброванных оценок L_k наиболее близких пользователей, а также полученную сумму разделим на сумму мер L выбранных пользователей. В итоге получим рекомендательный прогноз $R_{k,j}$ для j-го продукта относительно k-го пользователя. Данный алгоритм представим в виде формулы *3*:

$$\boldsymbol{R_{k,j} = k_{\text{норм}} \cdot \sum_{\substack{l=1,2,\dots,n \\ j=1,2,\dots,m \\ l \neq k}} m_{k,l} \cdot r_{k,j}}, \tag{3}$$

где нормировочный коэффициент $k_{\text{норм}}$ находим по формуле *4*:

$$\boldsymbol{k_{\text{норм}} = 1/\sum_{\substack{l=1,2,\dots,n \\ l \neq k}} m_{k,l}}. \tag{4}$$

Однако использование данного подхода в реализации метода коллаборативной фильтрации имеет свои недостатки. Во-первых, имея матрицу отношений размера $N \times M$, сложность подобного алгоритма можно оценить как $O(N{\cdot}M)$, что является проблемой при масштабировании информационной системы. Также данный алгоритм не позволяет в полной мере определить предпочтения конкретного пользователя.

Например, рассмотрим пару пользователей U_1 и U_2, состоящих в отношениях с объектами I_1 и I_2. Пусть U_1 оценил I_1, а U_2 оценил I_2, однако и U_1, и U_2 заинтересованы объектами I_1 и I_2. Очевидно, что при нахождении косинусной меры m_{U_1,U_2} полученный результат будет равен 0. Это означает, что объект I_1 не будет рекомендован пользователю U_2, а I_2 не будет рекомендован U_1.

Одним из возможных решений данной проблемы является применение алгоритма сингулярного разложения матриц (*SVD*).

Сингулярным разложением матрицы M порядка $n \times m$ является разложение [3], которое может быть записано в виде формулы *5*:

$$\boldsymbol{M = U \cdot \Sigma \cdot V^T}, \tag{5}$$

где матрица U порядка $n \times n$, V – размера $m \times m$, а Σ – размера $n \times m$.

Матрица Σ представляет собой диагональную матрицу, которая состоит из сингулярных чисел, расположенных в невозрастающем порядке. Таким образом, можно оставить лишь первые k сингулярных чисел и отбросить оставшиеся, получив наилучшее k-приближение исходной матрицы M (см. рисунок *1*).

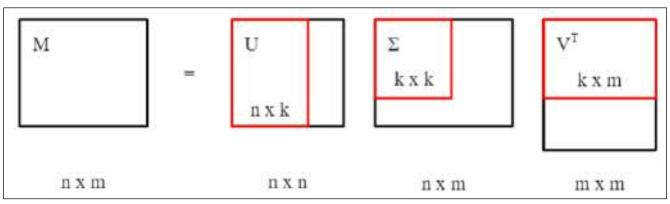


Рисунок 1 – Схематическое изображение *SVD*

Для построения Σ сперва получим матрицу A порядка $n \times n$ в результате произведения транспонированной матрицы на исходную по формуле *6*:

$$\boldsymbol{A} = \boldsymbol{M^T} \cdot \boldsymbol{M}, \tag{6}$$

откуда собственные значения матрицы A:

$$\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n,$$

где n – порядок матрицы A.

Тогда сингулярные значения A найдём путём вычисления квадратного корня из собственных значений по формулам *7*:

$$\boldsymbol{\sigma_1} = \sqrt{\boldsymbol{\lambda_1}}, \boldsymbol{\sigma_2} = \sqrt{\boldsymbol{\lambda_2}}, \boldsymbol{\sigma_3} = \sqrt{\boldsymbol{\lambda_3}}, \dots, \boldsymbol{\sigma_n} = \sqrt{\boldsymbol{\lambda_n}}, \tag{7}$$

Данные значения перенумеруем и расположим так, чтобы диагональные элементы матрицы Σ имели вид:

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \dots \geq \sigma_n > 0$$

Учтём то, что при перемножении исходной матрицы на транспонированную, полученная матрица B порядка $m \times m$ будет иметь такие же собственные значения, что и матрица A, однако они будут дополняться нулевыми значениями:

$$\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n, 0, \dots, 0.$$

Однако, при нахождении сингулярных значений, получим абсолютно идентичный набор, что содержит и матрица A:

$$\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_n.$$

Зная собственные значения матрицы A, можно найти её собственные векторы:

$$U_{\lambda=\lambda_1}; U_{\lambda=\lambda_2}; \dots; U_{\lambda=\lambda_n}.$$

Разделив каждый собственный вектор на его длину, получим нормированные векторы матрицы A, которые называются сингулярными:

$$\frac{U_{\lambda=\lambda_1}}{|U_{\lambda=\lambda_1}|}; \frac{U_{\lambda=\lambda_2}}{|U_{\lambda=\lambda_2}|}; \dots; \frac{U_{\lambda=\lambda_n}}{|U_{\lambda=\lambda_n}|}.$$

Данный набор сингулярных векторов образует матрицу U сингулярного разложения по столбцам.

Аналогичные рассуждения применимы к матрице B. Отличие заключается лишь в том, что набор сингулярных векторов матрицы B образует матрицу V^T сингулярного разложения (по строкам).

Важным свойством сингулярного разложения является тот факт, что если для k сингулярных чисел преобразовать матрицу Σ в матрицу:

$$\Sigma_k = \begin{array}{ccc} a_1 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & a_k \end{array} \in M(k,k),$$

состоящую только из k наибольших диагональных элементов, а также оставить в матрицах U и V^T лишь первые k столбцов и k строчек соответственно. Таким образом, матрицы U и V^Tбудут приведены к виду:

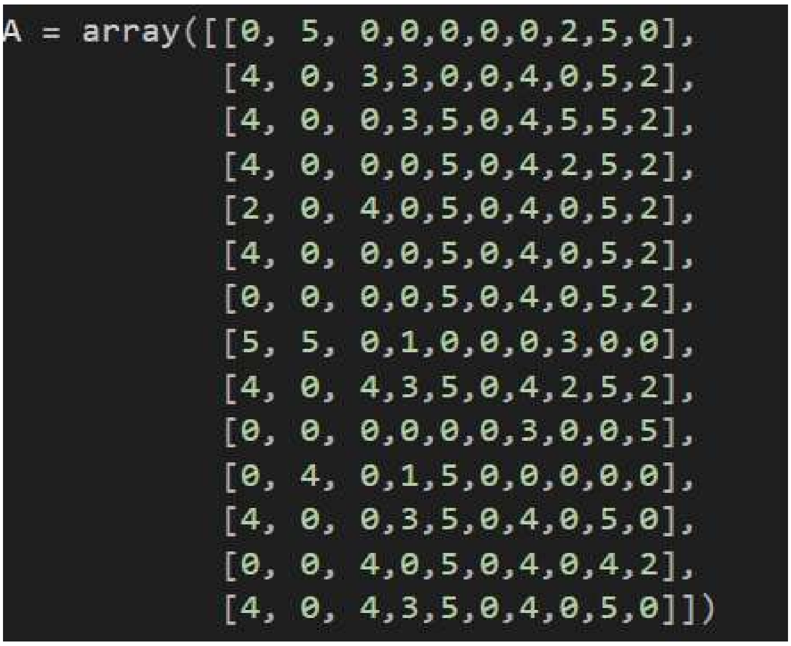
$$U_k \in M(n,k),$$
$$V^T{}_k \in M(k,m).$$

Данное усечение уменьшает размерность векторного пространства, тем самым, уменьшая объём данных для их обработки. Помимо этого, отбрасывая наименьшие сингулярные числа, малые искажения в результате шума в данных удаляются, оставляя только самые сильные эффекты и тенденции в этой модели. Полученный эффект улучшает качество предоставляемых рекомендаций.

Теперь для получения рейтинга r_{uv} пользователя u к объекту i необходимо лишь вычислить скалярное произведение i-ой строки матрицы U_k и j-ой строки матрицы $V^T{}_k$ по формуле 8:

$$r_{uv} = U_{k_i} \cdot V^T{}_{k_j}. \quad (8)$$

Смоделируем матрицу отношений размерностью 10 х 14. Исходный массив данных представлен на рисунке *2*.



```
A = array([[0, 5, 0,0,0,0,0,2,5,0],
           [4, 0, 3,3,0,0,4,0,5,2],
           [4, 0, 0,3,5,0,4,5,5,2],
           [4, 0, 0,0,5,0,4,2,5,2],
           [2, 0, 4,0,5,0,4,0,5,2],
           [4, 0, 0,0,5,0,4,0,5,2],
           [0, 0, 0,0,5,0,4,0,5,2],
           [5, 5, 0,1,0,0,0,3,0,0],
           [4, 0, 4,3,5,0,4,2,5,2],
           [0, 0, 0,0,0,0,3,0,0,5],
           [0, 4, 0,1,5,0,0,0,0,0],
           [4, 0, 0,3,5,0,4,0,5,0],
           [0, 0, 4,0,5,0,4,0,4,2],
           [4, 0, 4,3,5,0,4,0,5,0]])
```

Рисунок *2* – Исходная матрица отношений

В ходе программной реализации *SVD*, объём входных данных уменьшен, что явно делает систему более производительной, а также получена матрица рейтингов пользователей по отношению к объектам, по которой можно быстро определить данные значения (см. рисунок *3*).

```
[[-2.530e-01  5.753e-01 -4.743e-02 -6.250e-02 -2.294e-01 -2.070e-16 -9.753e-02  1.864e-01  6.245e-01 -6.470e-02]
 [ 2.347e-01  3.486e-02  2.003e-01  3.649e-01 -4.126e-01 -3.598e-16  2.548e-01 -2.128e-01  3.025e-01  1.552e-01]
 [ 2.272e-02 -1.879e-01 -1.840e-01  3.319e-01  2.084e-01  8.696e-16  8.731e-02  7.408e-01  1.294e-01  1.091e-01]
 [ 3.155e-01 -1.003e-01 -1.753e-01 -3.548e-01  2.181e-01  6.994e-16  1.274e-01  1.840e-01  1.902e-01  7.990e-02]
 [ 5.702e-02 -1.796e-03  4.314e-01 -3.138e-01  2.067e-01 -9.949e-17  1.141e-01 -4.463e-02  1.718e-01  7.665e-02]
 [ 3.740e-01 -1.646e-02 -2.335e-01 -3.227e-01  2.051e-01  2.822e-16  1.978e-01 -2.194e-01  1.931e-01  6.666e-02]
 [-2.594e-01  8.345e-03 -2.332e-01 -4.922e-02  2.579e-01  2.582e-16  1.898e-01 -1.024e-01  3.239e-01  1.427e-01]
 [ 6.047e-01  5.335e-01  5.769e-02 -8.120e-02 -6.754e-02  3.352e-16 -8.978e-02  2.330e-01 -2.050e-01  5.500e-02]
 [ 1.102e-01 -5.979e-02  3.935e-01  2.523e-01  1.642e-01 -1.259e-16  1.132e-01  2.520e-01  4.699e-02  6.123e-02]
 [-8.905e-02  1.175e-01 -9.486e-02  8.109e-02 -1.310e-01  1.537e-15  3.621e-01 -7.252e-02 -2.117e-01  7.740e-01]
 [-1.591e-01  5.435e-01 -6.071e-02  2.475e-01  5.512e-01 -3.776e-16 -4.227e-02 -1.514e-01 -2.645e-01  4.566e-03]
 [ 2.070e-01 -2.899e-02 -3.153e-01  4.060e-01  2.161e-01 -7.066e-16  1.645e-01 -2.808e-01  2.186e-01 -1.743e-01]
 [-2.270e-01  1.428e-02  4.532e-01 -1.582e-01  2.793e-01 -2.038e-17  1.123e-01  1.530e-02  1.080e-01  1.557e-01]
 [ 2.067e-01 -2.673e-02  3.494e-01  2.781e-01  1.913e-01 -1.076e-15  8.476e-02 -1.645e-01  1.319e-01 -2.023e-01]]
```

Рисунок *3* –Матрица рейтингов

Заключение. Алгоритм сингулярного разложения матриц позволяет оптимизировать и упростить классический метод коллаборативной фильтрации при построении рекомендательных систем. После применения алгоритма *SVD* к исходной матрице данные, имеющие несущественное влияние на рекомендацию или прогноз, не учитываются, а сам метод коллаборативной фильтрации упрощается от алгоритма из нескольких действий к одному (вычисление скалярного произведения), что увеличивает скорость построения рекомендаций и прогнозов.

Список литературы

1. Пятикоп, Е. Е. Использование сингулярного разложения матриц в коллаборативной фильтрации / Е. Е. Пятикоп // Проблеми інформатизації та управління.– 2013. – № 4(44) – С. 76 – 81. – Режим доступа: https://docplayer.com/47869869-Ispolzovanie-singulyarnogo-razlozheniya-matric-v-kollaborativnoy-filtracii.html – Дата доступа : 30.03.2022.

2. Количественный анализ и исследование эффективности методов поиска сходства в многомерных пространствах: материалы 24 Междунар. конф. VLDB, Нью-Йорк, 1998 / под ред. Вебер, Р., Шек, Х. Дж., Блотт, С. [и др.].– Нью-Йорк, 1998. – 194 с.

3. Логинов, Н.В. Сингулярное разложение матриц : учеб. пособие для вузов/ Н. В. Логинов. – Москва : МГАПИ, 1996. – 80 с.

UDC 004.78

THE ALGORITHM OF SINGULAR VALUE DECOMPOSITION IN THE COLLABORATIVE FILTERING METHOD FOR CONSTRUCTING A SYSTEM OF RECOMMENDATIONS AND FORECASTING

Romanovskiy N.D., Feoktistov D.G.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Rolich O.Ch. – PhD, assistant professor, associate professor of the department of ICSD

Annotation. The algorithm of singular value decomposition as one of the main ways of optimization of the collaborative filtering method and analysis of its impact on input data are considered in this work.

Keywords: recommendation systems, matrix decomposition, collaborative filtering