*УДК 621.45.01 : 004.945*

# СРАВНЕНИЕ АЛГОРИТМОВ ПОСТРОЕНИЯ АССОЦИАТИВНЫХ ПРАВИЛ НА ОСНОВЕ НАБОРА ДАННЫХ ПОКУПАТЕЛЬСКИХ ТРАНЗАКЦИЙ

© 2018 И.А. Олянич

Самарский национальный исследовательский университет имени академика С.П. Королев

Статья поступила в редакцию 12.12.2018

В статье рассмотрены алгоритмы построения ассоциативных правил Apriori и Eclat, с помощью которых производится анализ набора данных, содержащего в себе информацию о продуктовых покупках пользователей крупнейшего ритейлера в США Walmart. В ходе работы удается получить тривиальные и полезные правила, которые можно учитывать при формировании отделов магазина и расположении товаров таким образом, чтобы повысить покупательскую активность. Полученные графики позволяют визуально оценить построенные правила и сделать максимально точные прогнозы. Помимо этого, в статье выполнено сравнение двух алгоритмов нахождения ассоциативных правил по таким параметрам как изменение значения уровня поддержки и подача разного объема данных на вход.

*Ключевые слова:* Анализ данных, правило ассоциации, алгоритм Apriori, алгоритм Eclat, язык программирования R, RStudio, анализ рыночной корзины

## ВВЕДЕНИЕ

Анализ покупательской корзины относится к задачам интеллектуального анализа данных (data mining).

Data mining – это процесс поиска в большом объеме данных каких-либо закономерностей и получения знаний, которые требуются для принятия решений во многих сферах человеческой деятельности [1-4].

Правило ассоциации (associate rule) состоит из двух частей, предшествующей (если) и последующей (то). Предшествующая задача – это элемент, находящийся в данных. А последующая –это элемент или множество элементов, которые встречаются в сочетании с предшествующей задачей [5].

В интеллектуальном анализе данных правила ассоциации являются полезными и помогают спрогнозировать поведение клиента. Они играют важную роль в анализе покупательских корзин [6, 7].

Для оценки качества полученных рекомендаций используются следующие метрики:

1. *Поддержка* (support) позволяет узнать, в какой части покупательских корзин содержатся все элементы того или иного ассоциативного правила.

$$support(A \to B) = P(A \cup B).$$

2. *Достоверность* (confidence) показывает, насколько хорошим является правило для предсказания правой части, когда условие слева верно.

$$confidence(A \to B) = \frac{P(A \cup B)}{P(A)}.$$

*Олянич Игорь Анатольевич, аспирант.*
*E-mail: 14124123@mail.ru*

3. *Интерес* (lift) играет важную роль при анализе полученных правил и показывает, насколько хорошо было предсказано то, что находится в правой части. Другими словами, lift измеряет силу правила, сравнивая полное правило с предположенной правой частью и рассчитывается, как отношение достоверности правила к частоте появления следствия.

$$lift(A \to B) = \frac{P(A \cup B)}{P(A)P(B)}.$$

**Алгоритм Apriori**

Алгоритм Apriori использует горизонтальное представление множеств:

$$S_1 = \{a_{1,1}, a_{1,2}, \dots, a_{1,n_1}\},$$
$$S_2 = \{a_{2,1}, a_{2,2}, \dots, a_{2,n_2}\},$$
$$\dots$$
$$S_k = \{a_{k,1}, a_{k,2}, \dots, a_{k,n_k}\}.$$

$Contex$ – набор данных, $\min\_supp$ – минимальная поддержка,

$I_F$ – *все частые множества признаков.*

$C_1 \leftarrow \{1 - itemsets\}$

$i \leftarrow 1$

$while\ (C_i \neq 0)$

$$do \begin{cases} SupportCount(C_i) \\ F_i \leftarrow \{f \in C_i | f.support \geq \min\_supp\} // F - \text{частные множества признаков} \\ C_{i+1} \leftarrow AprioriGen(F_i) // C - \text{кандидаты} \\ i++ \end{cases}$$

$I_F \leftarrow \cup F_i$

$return(I_F)$

**Алгоритм Eclat**

На первом этапе алгоритм Eclat выполняет преобразование горизонтального предоставления множеств в вертикальное (по-другому можно назвать TID-множества) и в дальнейшем работа ведется именно с ним.

$$a_1 = \{S_{1,1}, S_{1,2}, \dots, S_{1,m_1}\},$$
$$a_2 = \{S_{2,1}, S_{2,2}, \dots, S_{2,m_2}\},$$
$$\dots$$
$$a_k = \{S_{k,1}, S_{k,2}, \dots, S_{k,m_k}\}.$$

В данном представлении поддержка будет выражаться, как отношение мощности множества к общему числу корзин.

$$\sup(A) = \frac{|A|}{N}.$$

Последующие этапы этого алгоритма аналогичны этапам алгоритма Apriori, кроме функции подсчета поддержки кандидата, которая теперь не требует сканирования базы.

## ИССЛЕДОВАТЕЛЬСКАЯ ЧАСТЬ

После реализации данных алгоритмов и оптимизации под исходный набор данных, были заданы уровень поддержки в диапазоне 0,001 и достоверность 0,8. Данные параметры можно считать оптимальными для получения наиболее полезных правил. Результат выполнения программы представлен в таблице 1.

Из полученной таблицы стоит выделить первое правило с высоким коэффициентом поддержки и интереса, что с первого взгляда может говорить о его полезности. Более подробно правило можно описать так: люди, покупающие ликер и вино с вероятностью 90% также приобретут пиво. Поддержка же говорит о том, что данные товары встречаются в 0,19% из общего числа транзакций, а интерес указывает на силу правила. Однако следует учесть, что товары находятся в одном и том же отделе магазина и наиболее вероятно расположены рядом друг с другом, поэтому полученное правило правильнее трактовать, как очевидное, нежели полезное. Последующие четыре правила тоже являются вполне очевидными, т.к. составляют средне статическую покупательскую корзину.

Чтобы выявить полезные правила, были предприняты попытки зафиксировать левую часть правила популярным продуктом и посмотреть полученные правила. Результат представлен в таблице 2.

Правила 1 и 5 могут считаться интересными, так как их следует учитывать при формировании отделов магазина. Если молочный отдел расположить рядом с кондитерскими изделиями, может увеличиться средний чек.

Чтобы оценить полученные правила, был построен график их разброса, представленный на рисунке 1.

Таким образом, удалось выявить, что наиболее оптимальным вариантов поиска полезных правил будет фиксирование в левой части нужного продукта и уже после анализ полученного результата. Данные действия могут помочь выстроить правильное расположение отделов, чтобы повысить покупательскую активность.

На заключительном шаге было произведено сравнение алгоритмов построения ассоциатив-

**Таблица 1.** Ассоциативные правила со значением supp = 0,001 и conf = 0,8

| № | lhs | | rhs | support | confidence | lift |
|---|---|---|---|---|---|---|
| 1 | {liquor,red/blush wine} | => | {bottled beer} | 0,0019 | 0,90 | 11,2 |
| 2 | {curd,cereals} | => | {whole milk} | 0,0010 | 0,91 | 3,6 |
| 3 | {yogurt,cereals} | => | {whole milk} | 0,0017 | 0,81 | 3,2 |
| 4 | {butter,jam} | => | {whole milk} | 0,0010 | 0,83 | 3,3 |
| 5 | {soups,bottled beer} | => | {whole milk } | 0,0011 | 0,92 | 3,6 |

**Таблица 2.** Ассоциативные правила для продукта «pastry»

| № | lhs | | rhs | support | confidence | lift |
|---|---|---|---|---|---|---|
| 1 | {pastry} | => | {whole milk} | 0,033 | 0,37 | 1,5 |
| 2 | {pastry} | => | {other vegetables} | 0,023 | 0,25 | 1,3 |
| 3 | {pastry} | => | {soda} | 0,021 | 0,24 | 1,4 |
| 4 | {pastry} | => | {rolls/buns} | 0,021 | 0,24 | 1,3 |
| 5 | {pastry} | => | {yogurt} | 0,018 | 0,20 | 1,4 |

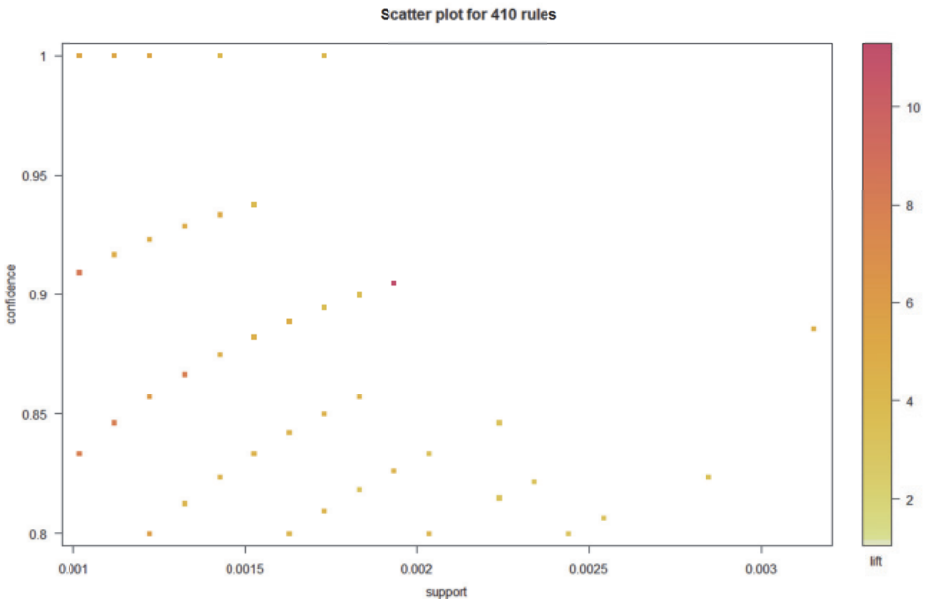


**Рисунок 1.** Разброс полученных правил

ных правил Apriori и Eclat в данной задаче для выбора наиболее оптимального.

Сначала изменялся параметр поддержки, результат можно увидеть в таблице 3, который свидетельствует о значительном преимуществе алгоритма Eclat, что в свою очередь можно обосновать работой с, так называемыми, TID-множествами.

На следующем шаге, параметр поддержки был фиксированный и подавалось разное количество данных на вход. Результат отображен в таблице 4.

Как можно заметить, алгоритм Apriori практически не способен работать на большом объеме данных. Таким образом, учитывая, что задачи анализа данных актуальны лишь при обработке большого числа входных значений, можно сделать вывод, что предпочтение стоит отдать алгоритму Eclat.

**Таблица 3.** Сравнение алгоритмов по значению поддержки

| Поддержка | Алгоритм Eclat | Алгоритм Apriori |
|---|---|---|
| 0,1 | 0,00 с | 0,00 с |
| 0,01 | 0,00 с | 0,01 с |
| 0,001 | 3,17 с | 5,01 с |
| 0,0001 | 5,25 с | 181,15 с |
| 0,00001 | 7,39 с | 427,37 с |

**Таблица 4.** Сравнение алгоритмов по объему данных

| Объем данных | Алгоритм Eclat | Алгоритм Apriori |
|---|---|---|
| 500 | 0,00 с | 0,00 с |
| 1 000 | 0,00 с | 0,01 с |
| 2 000 | 0,01 с | 0,22 с |
| 3 000 | 0,08 с | 0,92 с |
| 4 000 | 1,83 с | 102,98 с |

## СПИСОК ЛИТЕРАТУРЫ

1. *Наталия Е.* Введение в Data Mining [Электронный ресурс] // Компьютер пресс. 2016. URL: http://compress.ru/article.aspx?id=11616 (дата обращения 27.03.2018).
2. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP / Барсегян А.А. [и др.] - М. : БХВ-Петербург, 2007. - 384 с.
3. *Олянич И.А., Серафимович П.Г.* Сравнительное исследование алгоритмов проектирования рекомендательных систем на основе анализа крупноформатных данных о потребительских корзинах // Онтология проектирования, 2018, том 8, № 4(30), 628-640.
4. *Чубукова И.А.* Data Mining / И.А. Чубакова. - М. : Бином, 2008. - 324 с.
5. *Зайко Т.А., Олейник А.А., Субботин С.А.* Ассоциативные правила в интеллектуальном анализе данных [Электронный ресурс] // Киберленинка. URL: http://cyberleninka.ru/article/n/assotsiativnye-pravila-v-intellektualnom-analize-dannyh (дата обращения 17.05.2018).
6. *Шахиди А.* Data Mining — добыча данных [Электронный ресурс] // BaseGroupLabs Технологии анализа данных. 2016. URL: https://basegroup.ru/community/articles/data-mining (дата обращения 17.07.2018).
7. *Краковецкий А.* Анализ рыночной корзины и ассоциативные правила [Электронный ресурс] // Хабрхабр. URL: https://habrahabr.ru/post/66016/ (дата обращения 27.07.2018).


## COMPARISON OF ALGORITHMS OF CONSTRUCTION OF ASSOCIATIVE RULES ON THE BASIS OF THE DATA SET OF CUSTOMER TRANSACTIONS

© 2018 I.A. Olyanich

Samara National Research University named after Academician S.P. Korolyov

The article discusses the algorithms for constructing the association rules Apriori and Eclat, which are used to analyze a data set containing information about the grocery purchases of users of the largest US retailer Walmart. In the course of work, it is possible to obtain trivial and useful rules that can be taken into account when forming store departments and arranging goods in such a way as to increase consumer activity. The resulting graphs allow you to visually evaluate the constructed rules and make the most accurate predictions. In addition, the article compares two algorithms for finding associative rules for such parameters as changing the value of the support level and submitting a different amount of data to the input.

*Keywords:* Data mining, association rule, Apriori algorithm, Eclat algorithm, R programming language, RStudio, market basket analysis.

*Igor Olyanich, Postgraduate Student.*
*E-mail: 14124123@mail.ru*