МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
ФАКУЛЬТЕТ АВТОМАТИКИ И ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ

КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ

Отчет по лабораторной работе №2

по дисциплине «Методы анализа данных»

на тему «Поиск ассоциативных правил в среде Deductor»

Выполнили студенты группы АВТ-412:

Лазаревич М.М.

Евтушенко Н.С.

Проверила доцент кафедры АСУ:

Ганелина Н. Д.

г. Новосибирск, 2017

Цель работы:

Изучить процесс построения ассоциативных правил в программе *Deductor*.

Постановка задачи:

Провести поиск ассоциативных правил, проверить влияние параметров (поддержка, достоверность) на формирование правил.

Описание алгоритма:

Основным алгоритмом, который применяется для получения ассоциативных правил, является алгоритм apriori. Его автором является Ракеш Агравал.

Алгоритм Apriori предназначен для поиска всех частых множеств признаков. Он является поуровневым, использует стратегию поиска в ширину и осуществляет его снизу-вверх.

Алгоритм перебора следующий:

Алгоритм 2.5.1. APRIORI$(Context, min_supp)$

комментарий: Context - набор данных, min_supp - минимальная поддержка.

I_F — все частые множества признаков.

$C_1 \leftarrow \{1\text{-itemsets}\}$
$i \leftarrow 1$
while $(C_i \neq \emptyset)$
do $\begin{cases} SupportCount(C_i) \\ F_i \leftarrow \{f \in C_i \mid f.support \geq min\_supp\} // F - \text{частые множества признак} \\ C_{i+1} \leftarrow AprioriGen(F_i) // C - \text{кандидаты} \\ i{+}{+} \end{cases}$
$I_F \leftarrow \bigcup F_i$
return (I_F)

Алгоритм 2.5.2. AprioriGen(F_i)

комментарий: F_i - частые множества признаков длины i

C_{i+1} – потенциальные кандидаты частых множеств признаков.

insert into C_{i+1} // объединение
select $p[1], p[2], ..., p[i], q[i]$
from $F_i p, F_i q$
where $p[1] = q[1], ..., p[i-1] = q[i-1], p[i] < q[i]$
for each $c \in C_{i+1}$ // удаление
do {
  $S \leftarrow (i-1)$-элементные подмножества c
  for each $s \in S$
  do {
    if $(s \notin F_i)$
    then $C_{i+1} \leftarrow C_{i+1} \setminus c$
return (C_{i+1})

Основная особенность алгоритма — свойство антимонотонности.

Apriori использует одно из свойств поддержки, гласящее: поддержка любого набора элементов не может превышать минимальной поддержки любого из его подмножеств. Например, поддержка 3-элементного набора {Хлеб, Масло, Молоко} будет всегда меньше или равна поддержке 2-элементных наборов {Хлеб, Масло}, {Хлеб, Молоко}, {Масло, Молоко}. Дело в том, что любая транзакция, содержащая {Хлеб, Масло, Молоко}, также должна содержать {Хлеб, Масло}, {Хлеб, Молоко}, {Масло, Молоко}, причем обратное не верно.

Благодаря этому свойству перебор не является «жадным» и позволяет обрабатывать большие массивы информации за секунды.

Классический алгоритм apriori уже был несколько раз модифицирован, работы по улучшению скорости ведутся и сейчас.

Процедура применения метода:

Данный алгоритм был применён к массиву данных чеков из демонстрационного примера программы Deductor studio. Данный массив содержит 5000 записей вида Id Item, где Id – идентификатор чека, Item – наименование товара.

Применим алгоритм со следующими параметрами: поддержка 1-20, достоверность 25-40. В результате получается 46 правил. Правила характеризуются следующими показателями: Условие – основной набор товаров, вместе с которым с определённой частотой(достоверность) покупается другой набор товаров – следствие.

Поддержка – доля чеков, включающих в себя условие и следствие данного правила.
Лифт – отношение поддержки полного набора(условие и следствие) к произведению отдельных поддержек условия и следствия когда они встречаются отдельно друг от друга, достаточно важный показатель, позволяющий оценить полезность правила, если следствие встречается вне набора так же часто, как и в наборе, значит зависимость между условием и следствием скорее всего отсутствует, в таком случае лифт будет иметь значение 1, чем больше единицы значения лифта, тем существеннее связь между предметами, для ситуаций, когда лифт меньше нуля связь между предметами приобретает обратный характер, то есть для A->B можно сказать, что при lift(A->B)<1 товары B скорее всего не будут куплены вместе с товарами A.
Среди полученных результатов достаточно много неочевидных правил с большим значениям лифта, что свидетельствует о преобладании обоих совокупностей товаров вместе над их сочетаниями с другими товарами.

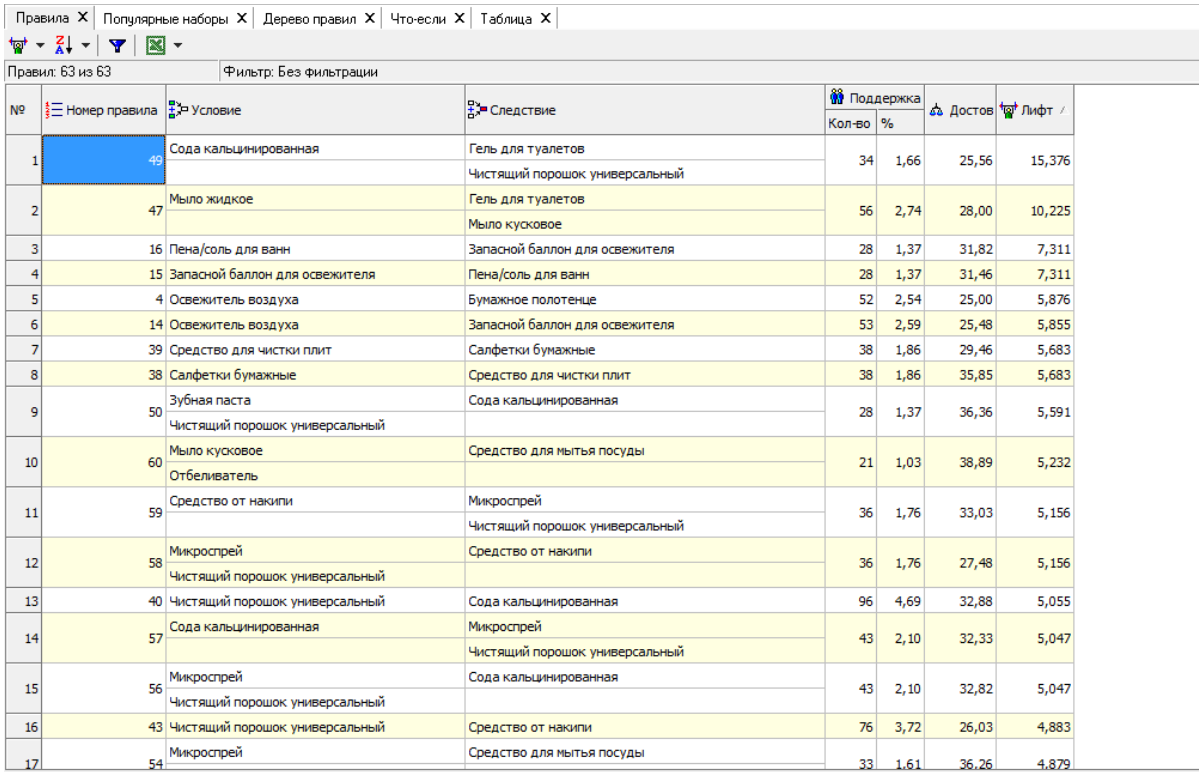
Правила | Популярные наборы | Дерево правил | Что-если | Таблица

Правил: 46 из 46 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достов	Лифт
				Кол-во	%		
1	40	Сода кальцинированная	Гель для туалетов	34	1,66	25,56	15,376
			Чистящий порошок универсальный				
2	38	Мыло жидкое	Гель для туалетов	56	2,74	28,00	10,225
			Мыло кусковое				
3	14	Пена/соль для ванн	Запасной баллон для освежителя	28	1,37	31,82	7,311
4	13	Запасной баллон для освежителя	Пена/соль для ванн	28	1,37	31,46	7,311
5	4	Освежитель воздуха	Бумажное полотенце	52	2,54	25,00	5,876
6	12	Освежитель воздуха	Запасной баллон для освежителя	53	2,59	25,48	5,855
7	30	Средство для чистки плит	Салфетки бумажные	38	1,86	29,46	5,683
8	29	Салфетки бумажные	Средство для чистки плит	38	1,86	35,85	5,683
9	41	Зубная паста	Сода кальцинированная	28	1,37	36,36	5,591
		Чистящий порошок универсальный					
10	43	Мыло кусковое	Средство для мытья посуды	21	1,03	38,89	5,232
		Отбеливатель					
11	31	Чистящий порошок универсальный	Сода кальцинированная	96	4,69	32,88	5,055
12	34	Чистящий порошок универсальный	Средство от накипи	76	3,72	26,03	4,883
13	33	Средство для чистки плит	Средство для мытья посуды	44	2,15	34,11	4,589
14	32	Средство для мытья посуды	Средство для чистки плит	44	2,15	28,95	4,589
15	25	Мыло кусковое	Средство для мытья посуды	131	6,41	33,42	4,496
16	46	Средство для чистки плит	Мыло кусковое	37	1,81	28,68	4,477
			Средство для мытья посуды				
17	45	Мыло кусковое	Средство для чистки плит	37	1,81	28,24	4,477
		Средство для мытья посуды					
18	35	Антистатик спрей	Мыло кусковое	28	1,37	27,72	4,328
			Средство для мытья посуды				

Рисунок 1. Правила..

В целях найти большее количество правил максимум поддержки был поднят до 40. Алгоритм был применён со следующими параметрами: поддержка 1-40, достоверность 25-40.
Как и следовало ожидать, количество выделенных правил не уменьшилось и даже увеличилось.



Правила | Популярные наборы | Дерево правил | Что-если | Таблица

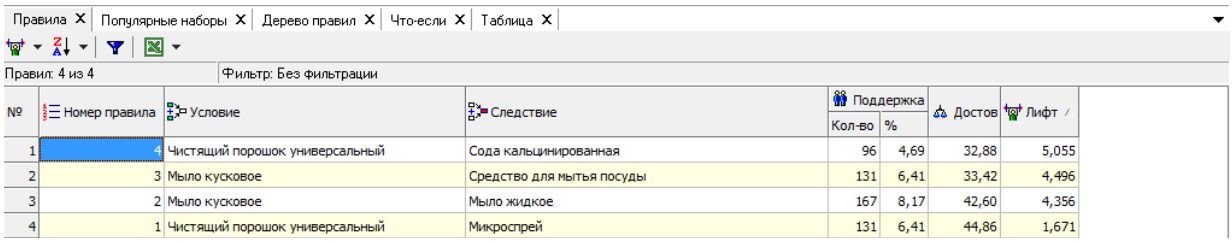
Правил: 63 из 63 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достов	Лифт
				Кол-во	%		
1	49	Сода кальцинированная	Гель для туалетов	34	1,66	25,56	15,376
			Чистящий порошок универсальный				
2	47	Мыло жидкое	Гель для туалетов	56	2,74	28,00	10,225
			Мыло кусковое				
3	16	Пена/соль для ванн	Запасной баллон для освежителя	28	1,37	31,82	7,311
4	15	Запасной баллон для освежителя	Пена/соль для ванн	28	1,37	31,46	7,311
5	4	Освежитель воздуха	Бумажное полотенце	52	2,54	25,00	5,876
6	14	Освежитель воздуха	Запасной баллон для освежителя	53	2,59	25,48	5,855
7	39	Средство для чистки плит	Салфетки бумажные	38	1,86	29,46	5,683
8	38	Салфетки бумажные	Средство для чистки плит	38	1,86	35,85	5,683
9	50	Зубная паста	Сода кальцинированная	28	1,37	36,36	5,591
		Чистящий порошок универсальный					
10	60	Мыло кусковое	Средство для мытья посуды	21	1,03	38,89	5,232
		Отбеливатель					
11	59	Средство от накипи	Микроспрей	36	1,76	33,03	5,156
			Чистящий порошок универсальный				
12	58	Микроспрей	Средство от накипи	36	1,76	27,48	5,156
		Чистящий порошок универсальный					
13	40	Чистящий порошок универсальный	Сода кальцинированная	96	4,69	32,88	5,055
14	57	Сода кальцинированная	Микроспрей	43	2,10	32,33	5,047
			Чистящий порошок универсальный				
15	56	Микроспрей	Сода кальцинированная	43	2,10	32,82	5,047
		Чистящий порошок универсальный					
16	43	Чистящий порошок универсальный	Средство от накипи	76	3,72	26,03	4,883
17	54	Микроспрей	Средство для мытья посуды	33	1,61	36,26	4,879

Рисунок 2. Правила.

Так, например, появилось правило: если куплено средство от накипи, то с достоверностью 33,03% будет куплен микроспрей с чистящим порошком(универсальным).

Изменим границы поддержки и достоверности следующим образом: поддержка 4-30, достоверность 25-60.



Правила | Популярные наборы | Дерево правил | Что-если | Таблица

Правил: 4 из 4 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достов	Лифт
				Кол-во	%		
1	4	Чистящий порошок универсальный	Сода кальцинированная	96	4,69	32,88	5,055
2	3	Мыло кусковое	Средство для мытья посуды	131	6,41	33,42	4,496
3	2	Мыло кусковое	Мыло жидкое	167	8,17	42,60	4,356
4	1	Чистящий порошок универсальный	Микроспрей	131	6,41	44,86	1,671

Рисунок 3. Правила.

С одной стороны, повышение минимальной поддержки должно уменьшить число не обоснованных правил, хотя при слишком большом значении пропадут и другие правила, повышение максимальной границы достоверности увеличивает количество тривиальных правил, то есть те случаи, когда причина связи между товарами вполне понятна и нахождение этих правил бесполезно, что и подтверждается результатами применения алгоритма. Из 4 полученных правил интерес представляет только последнее: если куплен чистящий порошок(универсальной), то с достоверностью 44,86% также будет куплен микроспрей.

Попробуем теперь выделить правила, представляющие интерес, применив алгоритм со следующими параметрами: поддержка 2-30, достоверность 25-30.

Правил: 16 из 16 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достов	Лифт
				Кол-во	%		
1	16	Мыло жидкое	Гель для туалетов	56	2,74	28,00	10,225
			Мыло кусковое				
2	1	Освежитель воздуха	Бумажное полотенце	52	2,54	25,00	5,876
3	6	Освежитель воздуха	Запасной баллон для освежителя	53	2,59	25,48	5,855
4	14	Чистящий порошок универсальный	Средство от накипи	76	3,72	26,03	4,883
5	13	Средство для мытья посуды	Средство для чистки плит	44	2,15	28,95	4,589
6	5	Гель для туалетов	Стиральный порошок ручной	66	3,23	29,86	3,132
7	15	Гель для туалетов	Мыло жидкое	56	2,74	25,34	3,103
			Мыло кусковое				
8	3	Гель для туалетов	Мыло жидкое	66	3,23	29,86	3,054
9	12	Пятновыводитель	Отбеливатель	41	2,00	28,08	2,886
10	8	Чистящий порошок универсальный	Зубная паста	77	3,77	26,37	1,872
11	7	Зубная паста	Чистящий порошок универсальный	77	3,77	26,74	1,872
12	11	Отбеливатель	Мыло кусковое	54	2,64	27,14	1,416
13	4	Гель для туалетов	Мыло кусковое	56	2,74	25,34	1,322
14	9	Средство для ухода за мебелью	Микроспрей	46	2,25	29,87	1,113
15	10	Стиральный порошок-автомат	Микроспрей	42	2,05	29,17	1,086
16	2	Гель для туалетов	Микроспрей	62	3,03	28,05	1,045

Рисунок 4. Правила.

Как видно, количество правил не так велико, но они могут быть с пользой использованы магазинами.

Вывод:

Проанализировав результаты, мы обнаружили, что для выделения наиболее полезных правил следует стремится к повышению минимума поддержки и понижению максимума достоверности, а обоснованием правдоподобности правила служит величина лифта.