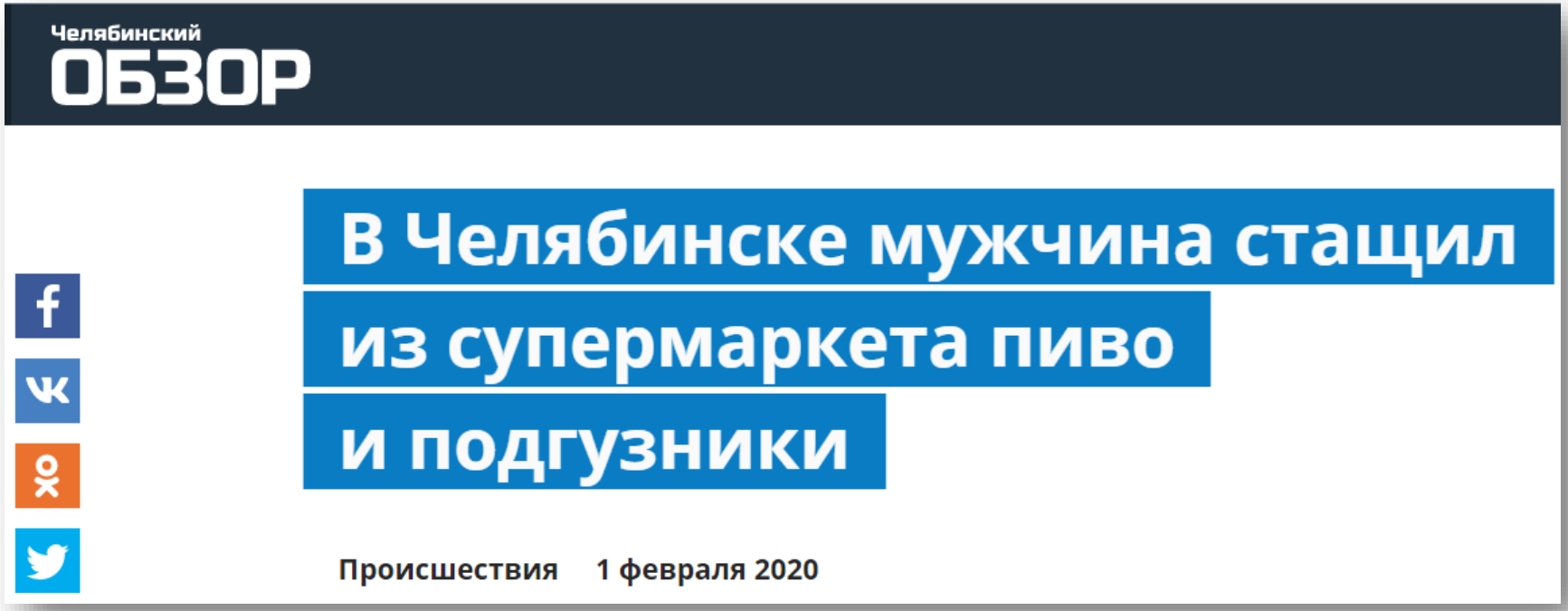
Задача поиска шаблонов в данных

Мыслить по шаблону –
вернейший способ завалить дело.

Дж.Ф. Энрайт


Челябинский ОБЗОР

В Челябинске мужчина стащил из супермаркета пиво и подгузники

Происшествия 1 февраля 2020

https://obzor174.ru/v-chelyabinske-muzhchina-stashchil-iz-supermarketa-pivo-i-podguzniki

Содержание

- Постановка задачи
  - Частые наборы
  - Ассоциативные правила
- Основные алгоритмы поиска частых наборов
  - Apriori
  - Eclat
  - FP-Growth
- **Меры полезности шаблонов:**
  - **support**
  - **confidence**
  - **lift и др.**
- Поиск шаблонов в Больших данных
  - Компактное представление частых наборов
  - Иерархии наборов
  - Фрагментация и сэмплинг для поиска частых наборов

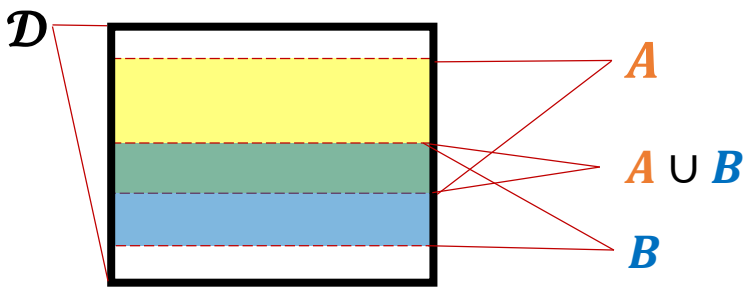
Шаблон, поддержка, достоверность

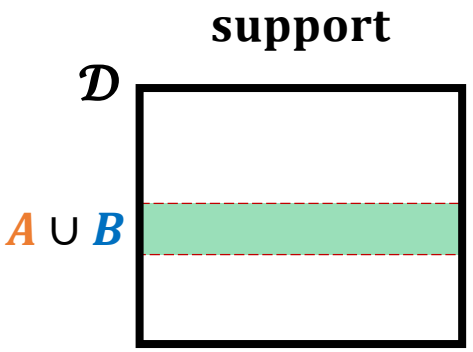
- Шаблон: $A \to B, A \neq \emptyset, B \neq \emptyset, A \cap B = \emptyset$
- Поддержка

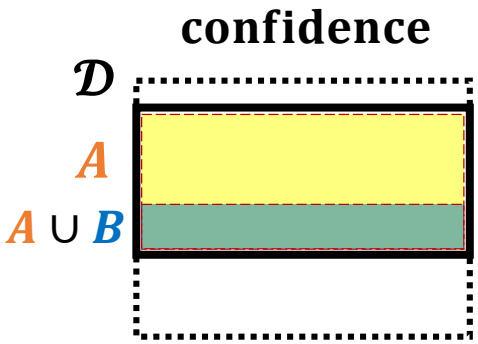
$$\sup(A \to B) = \mathrm{P}(A \cup B) = \frac{|\{t \in \mathcal{D} \mid (\mathrm{A} \cup \mathrm{B}) \subseteq t.I\}|}{|\mathcal{D}|}$$

- Достоверность

$$\mathrm{conf}(A \to B) = \mathrm{P}(B|A) = \frac{\sup(A \cup B)}{\sup(A)}$$







Полезность шаблонов

- Какие шаблоны полезны на практике?

№	Правило	sup	conf	Польза
1	хлеб → масло	👍	👍	?
2	хлеб → авокадо	👍	👍	?
3	молоко → тунец	👍	👎	?
4	масло → стир. порошок	👎	👎	?
5	водка → икра красная	👎	👍	?
6	водка → икра баклажанная	👎	👍	?

Полезность шаблонов

- Какие шаблоны полезны на практике?

№	Правило	sup	conf	Польза
1	хлеб → масло	👍	👍	👎
2	хлеб → авокадо	👍	👍	👍
3	молоко → тунец	👍	👎	👎
4	масло → стир. порошок	👎	👎	👎
5	водка → икра красная	👎	👍	👎
6	водка → икра баклажанная	👎	👍	👍

Пример: шаблон устойчив, но бесполезен

- Таблица сопряженности

	$coffee$	$\overline{coffee}$	Σ
tea	4000	3500	7500
$\overline{tea}$	2000	500	2500
Σ	6000	4000	10000

- $coffee \rightarrow tea\ [sup = \frac{4000}{10000} = 0.4, conf = \frac{4000}{6000} = 0.66]$
- $\mathrm{P}(tea) = \frac{7500}{10000} = 0.75 > \mathrm{P}(tea|coffee) = conf$
- Проще угадать наличие tea, чем предсказать это с помощью шаблона $coffee \rightarrow tea$

Пример: шаблон неустойчив, но полезен

- Таблица сопряженности

	$\boldsymbol{honey}$	$\overline{\boldsymbol{honey}}$	Σ
$\boldsymbol{tea}$	100	100	200
$\overline{\boldsymbol{tea}}$	20	780	800
Σ	120	880	1000

- $tea \rightarrow honey\ [sup = \frac{100}{1000} = 0.1, conf = \frac{100}{200} = 0.5]$
- $\mathrm{P}(honey) = \frac{120}{1000} = 0.12 < \mathrm{P}(honey|tea) = conf$
- $\mathrm{P}(\overline{tea} \cup honey) = \frac{20}{1000} = 0.025$
- Отбрасывание шаблона не учитывает информацию о предпочтении $honey$ при условии tea

Меры support и confidence не вполне адекватны

- Нужны дополнительные меры, которые учитывают корреляцию между объектами

Мера lift

- $\text{lift}(A, B) = \frac{P(A \cup B)}{P(A) \cdot P(B)}$
- Оценка
  - $\text{lift}(A, B) = 1$: A и B независимы
  - $\text{lift}(A, B) < 1$: A и B имеют отрицательную корреляцию (наличие A подразумевает отсутствие B)
  - $\text{lift}(A, B) > 1$: A и B имеют положительную корреляцию (наличие A подразумевает наличие B)
- $\text{lift}(A \to B) = \frac{P(B|A)}{P(B)} = \frac{\text{conf}(A \to B)}{\text{sup}(B)} = \frac{\text{sup}(A \to B)}{\text{sup}(A) \cdot \text{sup}(B)}$

Мера lift

- Пример:
  - Таблица сопряженности

	A	$\bar{A}$	Σ
B	4000	3500	7500
$\bar{B}$	2000	500	2500
Σ	6000	4000	10000

  - $A \rightarrow B$ $[sup = 0.4, conf = 0.66]$
- $\text{lift}(A, B) = \frac{P(A \cup B)}{P(A) \cdot P(B)} = \frac{0.4}{0.6 \cdot 0.75} = 0.89$
  - A и B имеют отрицательную корреляцию
  - $A \rightarrow B$ малополезно

Мера χ^2

- Выборки: $A = (a_1, \dots, a_c)$, $B = (b_1, \dots, b_r)$
- Таблица сопряженности $CT \in \mathbb{R}^{(r+1)\times(c+1)}$:
  - ячейки $(A = a_i, B = b_j)$, столбец и строка суммарных итогов
- $\chi^2 = \sum \frac{(observed-expected)^2}{expected} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$
  - $e_{ij} = \frac{count(A=a_i)\cdot count(B=b_j)}{n}$
  - $n = CT(r+1, c+1)$
- Таблица критических значений для отбрасывания гипотезы
  - Кол-во степеней свободы: $k = (r-1)(c-1)$
  - Уровень значимости: $\alpha \in (0,1)$

$k \setminus \alpha$	0.100	0.050	0.025	0.020	0.010	0.005	0.001
1	2.706	3.841	5.024	5.412	6.635	7.879	10.828
2	4.605	5.991	7.378	7.824	9.210	10.597	13.816
3	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	7.779	9.488	11.143	11.668	13.277	14.860	18.467
5	9.236	11.070	12.833	13.388	15.086	16.750	20.515
6	10.645	12.592	14.449	15.033	16.812	18.548	22.458
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	13.362	15.507	17.535	18.168	20.090	21.955	26.124

Мера χ^2

- Таблица сопряженности (вычислены ожидаемые значения)

	A	$\overline{A}$	Σ
B	4000 (4500)	3500 (3000)	7500
$\overline{B}$	2000 (1500)	500 (1000)	2500
Σ	6000	4000	10000

  - $A \rightarrow B$ $[sup = 0.4, conf = 0.66]$
- $\chi^2 = \frac{(4000-4500)^2}{4500} + \frac{(3000-3500)^2}{3000} + \frac{(2000-1500)^2}{1500} + \frac{(500-1000)^2}{1000} = 555.6;$
  - $\chi^2 > 1, o(A,B) < e(A,B) \Rightarrow$ отрицательная корреляция A и B
  - $A \rightarrow B$ малополезно

Другие меры

- $\text{allconf}(A,B) = \min(P(A|B), P(B|A)) = \frac{\sup(A \cup B)}{\max(\sup(A), \sup(B))}$
- $\text{maxconf}(A,B) = \max\big(P(A|B), P(B|A)\big) = \max\left(\frac{\sup(A \cup B)}{\sup(B)}, \frac{\sup(A \cup B)}{\sup(A)}\right)$
- $\text{Kulc}(A,B) = \frac{1}{2}\big(P(A|B) + P(B|A)\big) = \frac{1}{2}\left(\frac{\sup(A \cup B)}{\sup(B)} + \frac{\sup(A \cup B)}{\sup(A)}\right)$
- $\text{cosine}(A,B) = \frac{P(A \cup B)}{\sqrt{P(A) \cdot P(B)}} = \sqrt{P(A|B) \cdot P(B|A)} = \frac{\sup(A \cup B)}{\sqrt{\sup(A) \cdot \sup(B)}}$
- $\text{coherence}(A,B) = \frac{\sup(A \cup B)}{\sup(A) + \sup(B) - \sup(A \cup B)}$
- Общие свойства
  - зависимость от $\sup(A), \sup(B), \sup(A \cup B)$ (т.е. от $P(A|B), P(B|A)$)
  - отсутствие зависимости от $|\mathcal{D}|$
  - диапазон значений: $0..1$, большее означает большую связь A и B

Станислав Кульчиньский (1895-1975)

Какая мера лучше?

- Таблица сопряженности

	$\boldsymbol{milk}$	$\overline{\boldsymbol{milk}}$	$\Sigma_{\boldsymbol{row}}$
$\boldsymbol{coffee}$	mc	$\bar{m}c$	c
$\overline{\boldsymbol{coffee}}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
$\Sigma_{\boldsymbol{col}}$	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	$\boldsymbol{mc}$	$\boldsymbol{\bar{m}c}$	$\boldsymbol{m\bar{c}}$	$\boldsymbol{\overline{mc}}$	$\boldsymbol{\chi^2}$	**lift**	**all conf**	**max conf**	**Kulc**	**cosine**
$\boldsymbol{D_1}$	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
$\boldsymbol{D_2}$	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
$\boldsymbol{D_3}$	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
$\boldsymbol{D_4}$	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

Какая мера лучше?

- Таблица сопряженности

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

Какая мера лучше?

- Таблица сопряженности

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	**100000**	**90557**	**9.26**	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	**100**	**0**	**1**	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

Какая мера лучше?

- Таблица сопряженности

	$\boldsymbol{milk}$	$\overline{\boldsymbol{milk}}$	Σ_{row}
$\boldsymbol{coffee}$	mc	$\bar{m}c$	c
$\overline{\boldsymbol{coffee}}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	**100000**	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	**100**	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	**100000**	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

Какая мера лучше?

- Таблица сопряженности

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

Какая мера лучше?

- Таблица сопряженности

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

  - χ^2 и lift показывают неадекватные результаты ввиду чувствительности к $\overline{mc}$
  - как правило, $\overline{mc}$ нестабильно и велико

Null-транзакции и null-инвариантность мер

- Таблица сопряженности

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

- Сравнение мер для (m, c)

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	$lift$	all conf	max conf	Kulc	cosine
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5

m, c связаны положительно

m, c связаны отрицательно

m, c нейтральны

- χ^2 и lift не являются null-инвариантами

Null-инвариантная мера Imbalance Ratio (IR)

- $\mathrm{IR}(A,B) = \dfrac{|\sup(A) - \sup(B)|}{\sup(A) + \sup(B) - \sup(A \cup B)}$

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\bar{m}c$	c
$\overline{coffee}$	$m\bar{c}$	$\overline{mc}$	$\bar{c}$
Σ_{col}	m	$\bar{m}$	Σ

$\mathcal{D}$	mc	$\bar{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	lift	all conf	max conf	Kulc	cosine	IR
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91	0
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91	0
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09	0
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	**0.5**	0.5	**0**
D_5	1000	10	10000	100000	8173	9.18	0.09	0.91	**0.5**	0.29	**0.89**
D_6	1000	10	100000	100000	965	1.97	0.01	0.99	**0.5**	0.1	**0.99**

Резюме о мерах полезности шаблонов

- Применение при поиске только мер sup и conf приводит к большому количеству шаблонов, многие из которых неадекватны
- Рекомендуется дополнительно использовать меры, обладающие свойством null-инвариантности
- Среди мер с этим свойством рекомендуется совместное использование Kulc и IR

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - Chapter 6. Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods, pp. 243-278
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN-13: 978-0-13-312890-1
  - 5. Association Analysis: Basic Concepts and Algorithms, pp. 357-450