

# Майнор “Введение в анализ данных”

## Частые множества и ассоциативные правила

Игнатов Дмитрий Игоревич<sup>◊</sup>

<sup>◊</sup>Национальный исследовательский университет Высшая школа экономики  
Факультет компьютерных наук  
Департамент анализа данных и искусственного интеллекта

AD-minor 2016

# Содержание

## 1 Основная часть

- Введение
- Области применения
- Анализ формальных понятий
- Частые множества и ассоциативные правила
- Алгоритм Apriori
- Алгоритм FP-growth
- Меры интересности правил
- Компактное представление частых множеств

## 2 Прикладные задачи и эксперименты

- Анализ посещаемости веб-сайтов
- Рекомендация контекстной рекламы

## 3 Программные средства

## 4 Чего бы почитать и посмотреть?

# Введение

## KDD & Data Mining

- Data mining основной этап при обнаружении знаний в базах данных (Knowledge Discovery in Databases)
- Поиск ассоциативных правил (association rules) и частых множеств признаков (frequent itemset mining) одни из ключевых методов Data Mining
- Исходная задача — анализ потребительской корзины

# О терминологии. KDD и Data Mining

## Knowledge discovery in Databases (KDD)

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Fayyad, Piatetsky-Shapiro, and Smyth 1996

## Data Mining

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

Там же

# О терминологии. KDD и Data Mining

## Схема процесса обнаружения знаний в данных

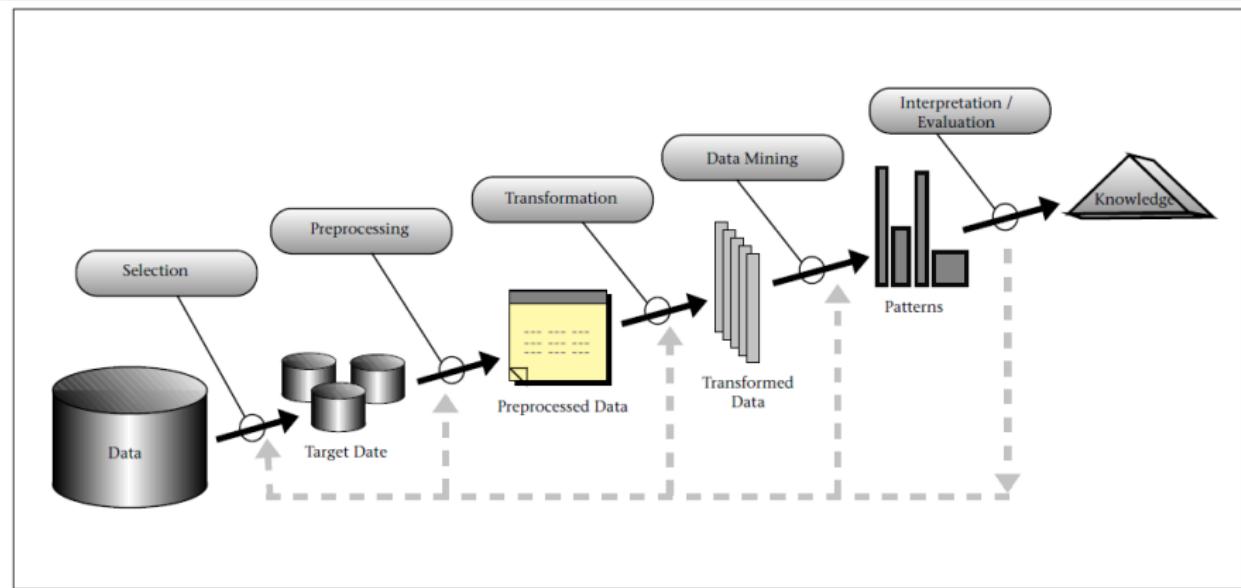


Figure 1. An Overview of the Steps That Compose the KDD Process.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

# О терминологии. KDD и Data Mining

[J. Han et al., Data Mining. Concepts and Techniques, 3rd Ed., 2012]

- ① Data cleaning
- ② Data integration
- ③ Data selection
- ④ Data transformation
- ⑤ Data mining (an essential process where intelligent methods are applied to extract data patterns)
- ⑥ Pattern evaluation
- ⑦ Knowledge presentation

## Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

# О терминологии. Машинное обучение

[T. Mitchell. The Discipline of Machine Learning, 2006]

## Основной вопрос в машинном обучении

How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

### Более точно

To be more precise, we say that a **machine learns** with respect to a particular task  $T$ , performance metric  $P$ , and type of experience  $E$ , if the system reliably improves its performance  $P$  at task  $T$ , following experience  $E$ . Depending on how we specify  $T$ ,  $P$ , and  $E$ , the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

# О межпредметных связях

## Гипотеза

Data Mining  $\stackrel{?}{=}$  Machine Learning

## Связанные дисциплины

- Computer Science (Информатика)
- Artificial Intelligence (Искусственный интеллект)
- Pattern Recognition (Распознавание образов)
- Information Retrieval (Информационный поиск)
- Social Network Analysis (Анализ социальных сетей)
- Теория вероятностей и математическая статистика
- Дискретная математика (в т.ч. порядки и графы)
- Optimization (Методы оптимизации)

# Области применения DM&ML

## Области применения

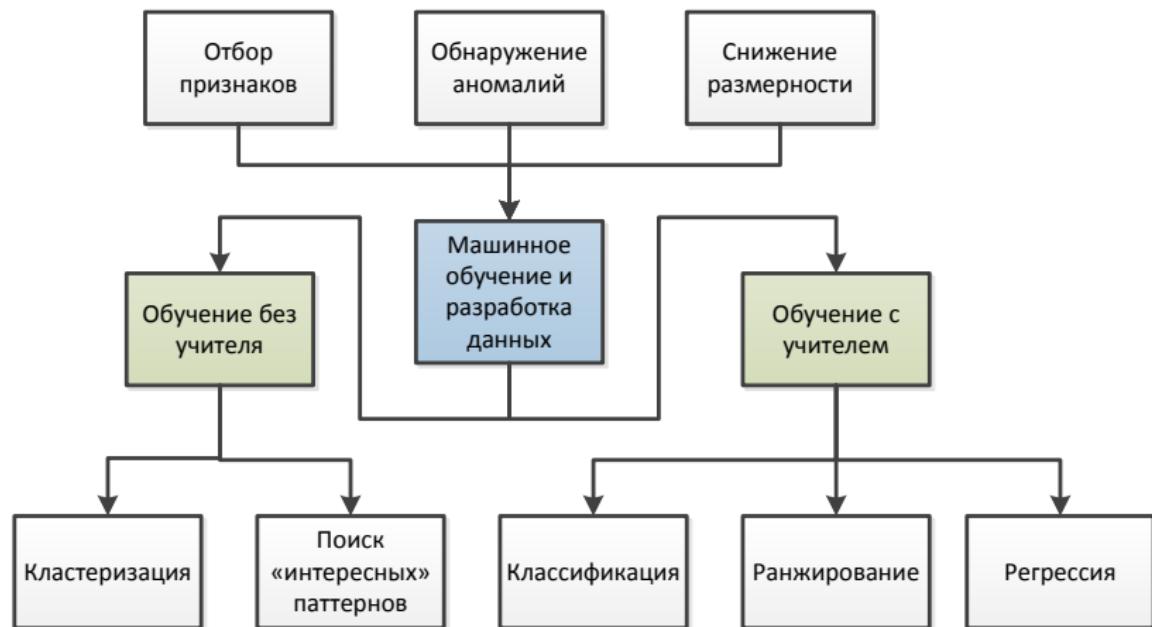
- Бизнес
- Медицина
- Образование
- Науки о жизни
- Интернет-данные
- Банковское дело и финансы
- ...

# Тренды в областях применения DM&ML

[J. Han et al., 2012]

- Application exploration: e.g., counter-terrorism and mobile (wireless) data mining
- Scalable and interactive data mining methods
- Integration of data mining with search engines, database systems, data warehouse systems, and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving-objects, and cyber-physical system
- Mining multimedia, text, and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

# Таксономия методов DM&ML



# Поиск паттернов/ зависимостей

## Постановка задачи

- Поиск закономерностей в данных об использовании каких-либо ресурсов, например, часто используемых вместе.
- Пример:  $support(\{\text{хлеб, молоко}\}) = 0.7$
- Часто такие закономерности записываются в виде правил

$$A \longrightarrow B$$

- Пример: {Студент, Возраст от 16 до 25}  $\longrightarrow \{iPhone, iPad\}$

# Поиск паттернов/зависимостей



# Анализ Формальных Понятий

[Wille, 1982], [Ganter, 1999]

- $G$  — множество **объектов**,  $M$  — множество **признаков**
- отношение  $I \subseteq G \times M$ , такое что  $gIm$ , тогда и только тогда, когда объект  $g$  обладает признаком  $m$ .
- $\mathbb{K} = (G, M, I)$  называется **формальным контекстом**.

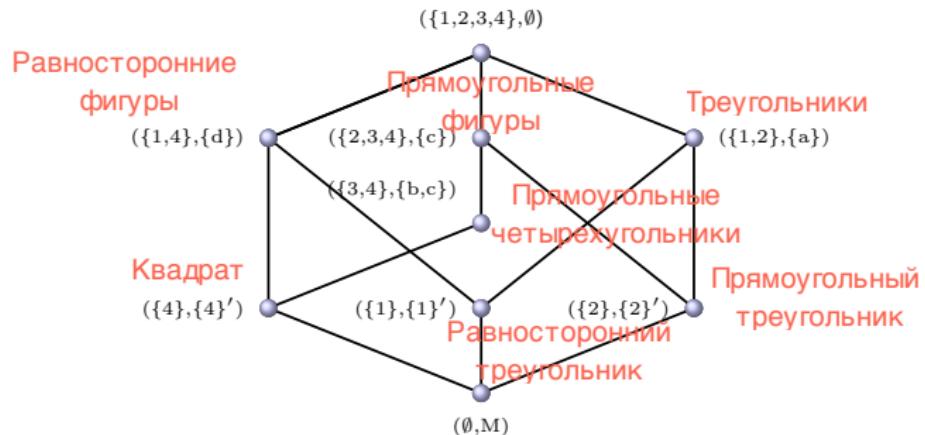
**Операторы Галуа:**  $A \subseteq G, B \subseteq M$

$$A' = \{m \in M \mid gIm \text{ для всех } g \in A\}, B' = \{g \in G \mid gIm \text{ для всех } m \in B\}.$$

**Формальное понятие** есть пара  $(A, B)$ :  $A \subseteq G, B \subseteq M, A' = B, B' = A$ .

- $A$  называется **(формальным) объемом**, а  $B$  называется **(формальным) содержанием** понятия  $(A, B)$ .
- Понятия, упорядоченные отношением  $(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2$ , образуют полную решетку, называемую **решеткой понятий**  $\underline{\mathfrak{B}}(G, M, I)$ .
- Оператор  $(\cdot)''$  является оператором замыкания (идемпотентен, монотонен, экстенсивен)

# АФП: пример контекста и решетки понятий



	$G \setminus M$	a	b	c	d
1		x			x
2		x		x	
3			x	x	
4		x	x	x	x

- a – ровно 3 вершины,
- b – ровно 4 вершины,
- c – имеет прямой угол,
- d – все стороны равны

# Импликации на подмножествах признаков

## Определение

**Импликация**  $A \rightarrow B$ , где  $A, B \subseteq M$ , имеет место если  $A' \subseteq B'$ , т.е. каждый объект, обладающий всеми признаками из множества  $A$ , также обладают всеми признаками из множества  $B$ .

## Определение

Импликации удовлетворяют **правилам Армстронга**:

$$\frac{}{X \rightarrow X}, \quad \frac{X \rightarrow Y}{X \cup Z \rightarrow Y}, \quad \frac{X \rightarrow Y, Y \cup Z \rightarrow W}{X \cup Z \rightarrow W}$$

# Основные определения

## Определение 1

Пусть дан контекст  $\mathbb{K} := (G, M, I)$ , где  $G$  — множество объектов (транзакций, покупок),  $M$  — множество признаков (товаров, items),  $I \subseteq G \times M$

Ассоциативным правилом контекста  $\mathbb{K}$  называется выражение вида  $A \rightarrow B$ , где  $A, B \subseteq M$ .

Часто требуют  $A \cap B = \emptyset$

# Основные определения

## Определение 2

Поддержкой (support) ассоциативного правила  $A \rightarrow B$  называется величина  $supp(A \rightarrow B) = \frac{|(A \cup B)'|}{|G|}$ .

Значение  $supp(A \rightarrow B)$  показывает какая доля объектов  $G$  содержит  $A \cup B$ . Часто поддержку выражают в %.

# Основные определения

## Определение 2

Поддержкой (support) ассоциативного правила  $A \rightarrow B$  называется величина  $supp(A \rightarrow B) = \frac{|(A \cup B)'|}{|G|}$ .

Значение  $supp(A \rightarrow B)$  показывает какая доля объектов  $G$  содержит  $A \cup B$ . Часто поддержку выражают в %.

## Определение 3

Достоверностью (confidence) ассоциативного правила  $A \rightarrow B$  называется величина  $conf(A \rightarrow B) = \frac{|(A \cup B)'|}{|A'|}$ .

Значение  $conf(A \rightarrow B)$  показывает какая доля объектов обладающих  $A$  также содержит  $A \cup B$ . Величину поддержки также часто выражают в %.

# Основные определения

## Определение 4

Множество признаков  $F \subseteq M$  называется **частым множеством признаков** если  $supp(F) \geq min\_supp$ .

# Пример

## Объектно-признаковая таблица транзакций

Покупатели/товары	Пиво	Пряники	Молоко	Мюсли	Чипсы
c <sub>1</sub>	1	0	0	0	1
c <sub>2</sub>	0	1	1	1	0
c <sub>3</sub>	1	0	1	1	1
c <sub>4</sub>	1	1	1	0	1
c <sub>5</sub>	0	1	1	1	1

- $supp(\{\text{Пиво}, \text{Чипсы}\}) = 3/5$
- $supp(\{\text{Пряники}, \text{Мюсли}\} \rightarrow \{\text{Молоко}\}) = \frac{|(\{\text{Пряники}, \text{Мюсли}\} \cup \{\text{Молоко}\})'|}{|G|} = \frac{|\{c_2, c_5\}|}{5} = 2/5$
- $conf(\{\text{Пряники}, \text{Мюсли}\} \rightarrow \{\text{Молоко}\}) = \frac{|(\{\text{Пряники}, \text{Мюсли}\} \cup \{\text{Молоко}\})'|}{|\{\text{Пряники}, \text{Мюсли}\}'|} = \frac{|\{c_2, c_5\}|}{|\{c_2, c_5\}'|} = 1$

# Постановка задачи

## Поиск ассоциативных правил, min-confidence и min-support

Требуется найти все ассоциативные правила контекста, для которых значения поддержки и достоверности превышают некоторые установленные значения,  $\text{min\_supp}$  и  $\text{min\_conf}$  соответственно [Agrawal et al., 1993].

## Ассоциативные правила и импликации

- Ассоциативные правила при значениях  $\text{min\_supp} = 0\%$  и  $\text{min\_conf} = 100\%$  являются импликациями рассматриваемого контекста.
- Иногда ассоциативные правила записывают в форме  $A \xrightarrow[s]{c} B$ , где  $s$  — confidence и  $c$  support данного правила соответственно.

# Поиск ассоциативных правил

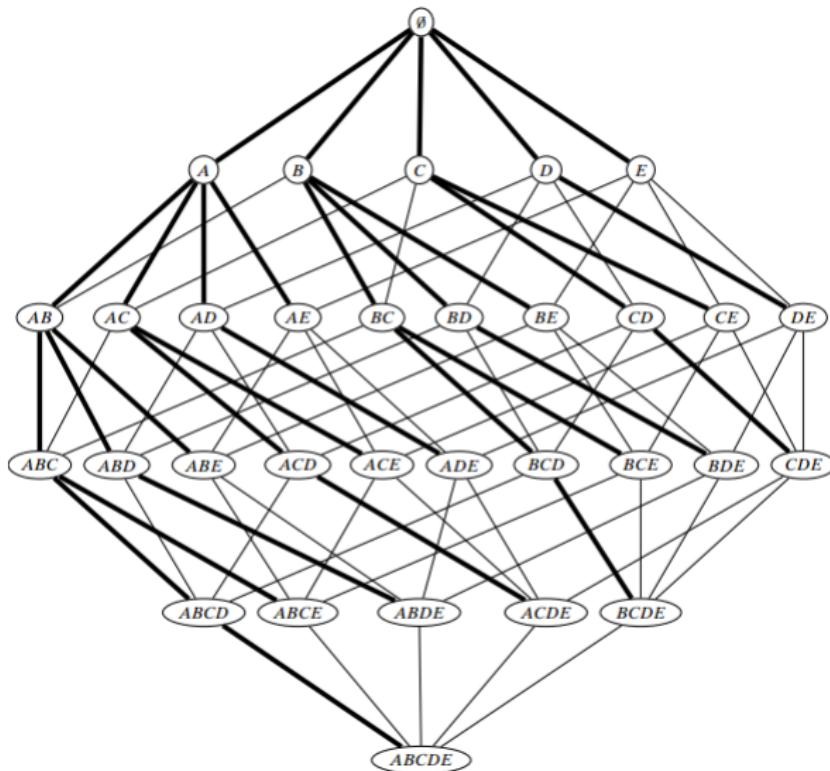
## Этапы поиска

- ① Нахождение частых множеств признаков (frequent itemsets), т.е. множеств признаков с поддержкой не ниже заданой ( $min\_supp$ ).
- ② Построение ассоциативных правил на основе найденных частых множеств признаков.

- Первый шаг наиболее трудоемкий, второй шаг тривиальный.
- Классический алгоритм, строящий частые множества признаков — Apriori [Agrawal, Srikant, 1994]

# Поиск частых множеств

Обход булевой решетки множеств



# АФП мирно встречает Data Mining

- Agrawal R., RSFDGrC – 2011, Москва



# Антимонотонность

## Свойство 1 (антимонотонность)

Для  $\forall A, B \subseteq M$  и  $A \subseteq B \Rightarrow supp(B) \leq supp(A)$

- Ключевое свойство при нахождении многоэлементных частых множеств признаков
- С ростом размера множества его поддержка уменьшается, либо не изменяется
- Поддержка любого множества признаков не превышает минимальной поддержки любого его подмножества
- Множество признаков размера  $n$  будет частым, когда все его  $(n - 1)$ -элементные подмножества будут частыми

# Алгоритм Apriori

## Описание

находит все частые множества признаков

---

### Алгоритм 1.1. Apriori(*Context*, *min\_supp*)

---

input: *Context* – набор данных, *min\_supp* – минимальная поддержка

output: все частые множества признаков  $I_F$

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

$i \leftarrow 1$

while ( $C_i \neq \emptyset$ )

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

$I_F \leftarrow \bigcup F_i$

return ( $I_F$ )

---

# Процедура AprioriGen

## Описание

для  $i$ -элементных частых множеств признаков порождает их  $(i + 1)$ -надмножества и возвращает только множество потенциально частых кандидатов

---

### Алгоритм 1.2. AprioriGen( $F_i$ )

---

input:  $F_i$  – частые множества признаков длины  $i$

output:  $C_{i+1}$  – потенциальные кандидаты частых множеств признаков

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

---

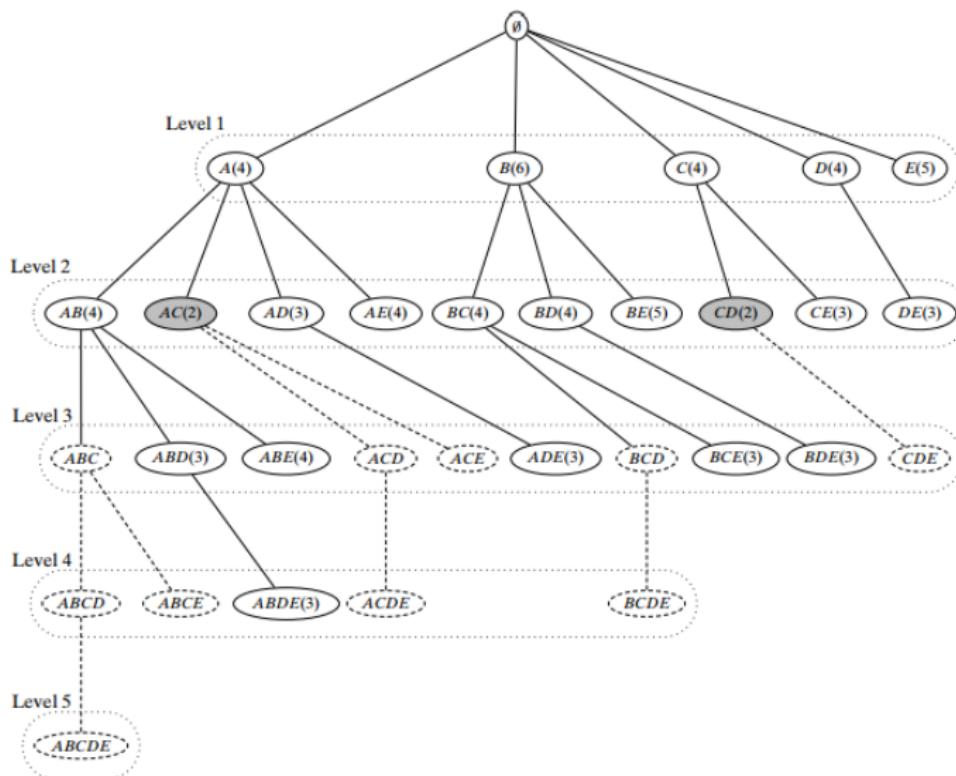
# Пример работы AprioriGen

## Шаги объединение и исключение

- $F_3 = \{\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}\}$
- $C_4 = \{\{a, b, c, d\}, \{a, c, d, e\}\}$  – шаг объединение
- $C_4 = \{\{a, b, c, d\}\}$ , исключаем  $\{a, c, d, e\}$ , т.к. его подмножество  $\{c, d, e\} \notin F_3$  – шаг удаление

# Поиск частых множеств

## Решетка частых множеств



# Построение правил

## Извлечение правил из частых множеств признаков

Пусть  $F$  — частое множество признаков. Записываем правило  $f \rightarrow F \setminus f$ , если

$$conf(f \rightarrow F \setminus f) = \frac{supp(F)}{supp(f)} \geq min\_conf$$

# Построение правил

## Свойство 2

$conf(f \rightarrow F \setminus f) = \frac{supp(F)}{supp(f)}$  имеет минимальное значение, когда  $s(f)$  максимально.

- Достоверность минимальна, когда посылка правила состоит из одного признака. Надмножества такого признака имеют меньшую поддержку, а значит, и большую достоверность.
- Рекурсивная процедура извлечения правил. Начинаем с одноэлементной посылки  $f$  удовлетворяющей  $min\_conf$  и  $min\_sup$ , проверяем все надмножества для данного  $F$ . Используем обязательно все признаки из  $F$  на каждом шаге для построения правила.

# Задание

- ① С помощью алгоритма Apriori построить все частые множества признаков контекста из примера 1 для значения  $min\_sup = 1/3$

# Задание

- ➊ С помощью алгоритма Apriori построить все частые множества признаков контекста из примера 1 для значения  $min\_sup = 1/3$
- ➋ Please, say “I ❤️ Apriori”.

# Алгоритм FP-growth

[Han et al., 2000]

- Jiawei Han, Jian Pei, Yiwen Yin: Mining Frequent Patterns without Candidate Generation. SIGMOD Conference 2000: 1-12
- Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min. Knowl. Discov. 8(1): 53-87 (2004)

# Алгоритм FP-growth

Данные для примера (Zaki & Meira, 2014)

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary database

t	i(t)
1	ABDE
2	BCE
3	ABDE
4	ABCE
5	ABCDE
6	BCD

(b) Transaction database

x	A	B	C	D	E
	1	1	2	1	1
	3	2	4	3	2
t(x)	4	3	5	5	3
	5	4	6	6	4
			5		
			6		

(c) Vertical database

# Алгоритм FP-growth

FP-дерево: транзакции 1-4



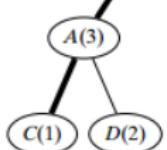
(a) (1, BEAD)



(b) (2, BEC)



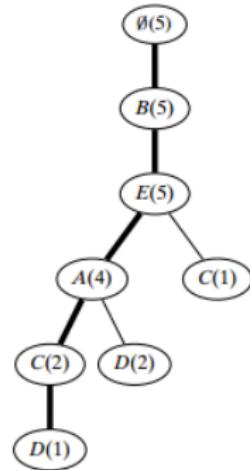
(c) (3, BEAD)



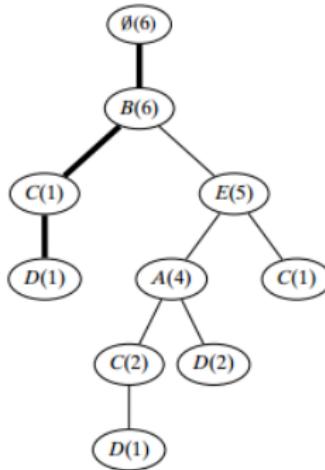
(d) (4, BEAC)

# Алгоритм FP-growth

FP-дерево: транзакции 5-6



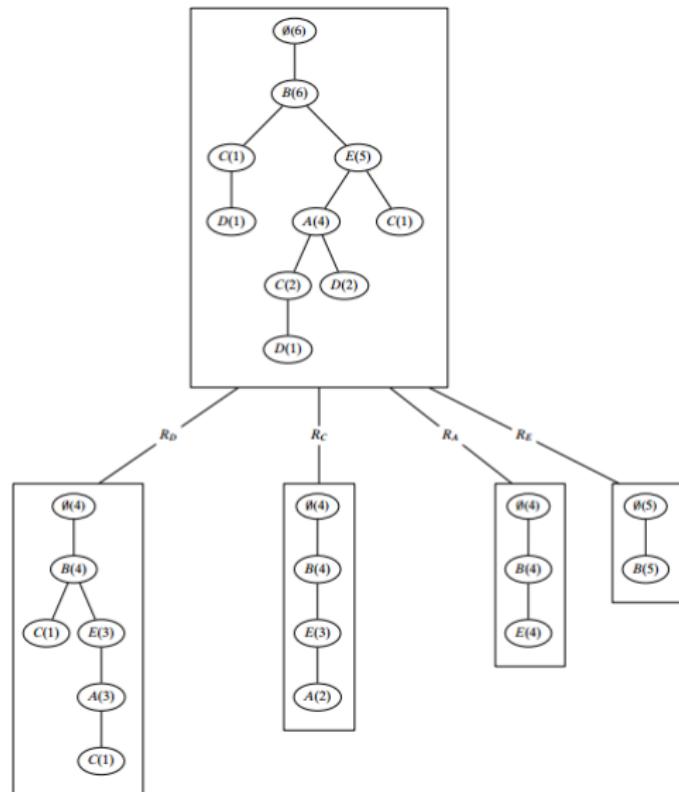
(e)  $\langle 5, BEACD \rangle$



(f)  $\langle 6, BCD \rangle$

# Решетка частых множеств

Проекция для D



# Меры интересности правил

Zaki & Meira 2014, Глава 12 “Pattern and Rule Assessment”

## Коэффициент Жаккара

$$Jaquard(A, B) = \frac{|A' \cap B'|}{|A' \cup B'|} = \frac{sup(AB)}{sup(A) + sup(B) - sup(AB)}$$

## “Поднятие” (lift) $A \rightarrow B$

$$lift(A, B) = \frac{P(AB)}{P(A)P(B)} = \frac{P(A|B)}{P(A)}$$

## “Поднятие” правила $\neg A \rightarrow B$

$$lift(\neg A, B) = \frac{P(\neg AB)}{P(\neg A)P(B)} = \frac{P(\neg A|B)}{P(\neg A)}$$

# Компактное представление частых множеств признаков

Пусть дан контекст  $\mathbb{K} := (G, M, I)$

## Определение 5

Множество признаков  $FC \subseteq M$  называется **частым замкнутым множеством признаков** если  $supp(FC) \geq min\_supp$  и не существует  $F$ , такого что  $F \supset FC$  и  $supp(F) = supp(FC)$ .

## Определение 6

Множество признаков  $MFC \subseteq M$  называется **максимальным частым замкнутым множеством признаков** если оно частое и не существует  $F$ , такого что  $F \supset MFC$  и  $supp(F) \geq min\_supp$ .

# Компактное представление частых множеств признаков

Пусть дан контекст  $\mathbb{K} := (G, M, I)$

## Утверждение 1

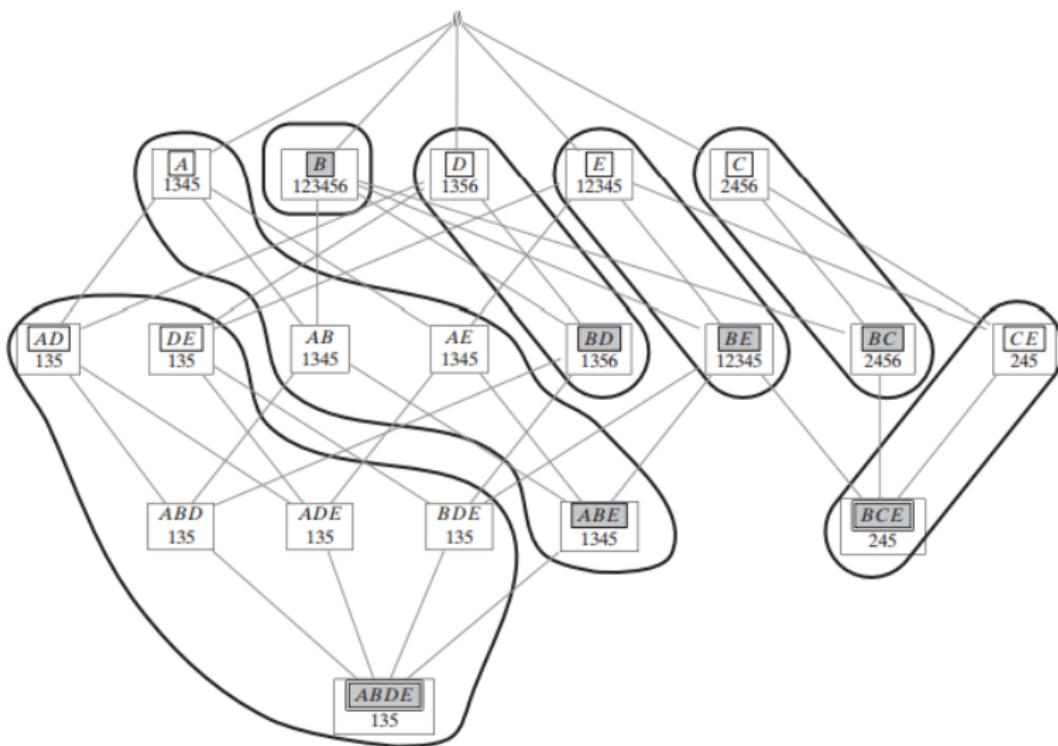
$\mathcal{MFC} \subseteq \mathcal{FC} \subseteq \mathcal{F}$ , где  $\mathcal{MFC}$  – максимальные замкнутые множества признаков контекста  $\mathbb{K}$ ,  $\mathcal{FC}$  – частые замкнутые множества признаков, а  $\mathcal{F}$  – частые множества признаков для заданной минимальной поддержки  $min\_supp$ .

## Утверждение 2

Решетка формальных понятий контекста  $\mathbb{K}$  изоморфна решетке его частых замкнутых множеств признаков для заданной минимальной поддержки  $min\_supp = 0$ .

# Решетка частых множеств

Максимальные и замкнутые множества



# Содержание

## 1 Основная часть

- Введение
- Области применения
- Анализ формальных понятий
- Частые множества и ассоциативные правила
- Алгоритм Apriori
- Алгоритм FP-growth
- Меры интересности правил
- Компактное представление частых множеств

## 2 Прикладные задачи и эксперименты

- Анализ посещаемости веб-сайтов
- Рекомендация контекстной рекламы

## 3 Программные средства

## 4 Чего бы почитать и посмотреть?

# Постановка задачи

ООО “Мастерхост”, 2006-2007

- По данным, собираемым счетчиками посещений на Интернет-сайтах, требуется выявлять интересы аудитории целевого сайта
- Предлагается модель построения таксономий пользователей (аудиторий) веб-сайтов на основе АФП с применением критериев отбора релевантных формальных понятий

# Математическая модель таксономий аудиторий веб-сайтов

## Внешняя таксономия

$\mathbb{K}_{ex} = (V, S_{ex}, I)$ , где

$V$  – множество всех посетителей целевого сайта,  $S_{ex}$  – множество всех сайтов выборки исключая целевой,  $I$  – отношение инцидентности  $vIs$ ,  $v \in V$ ,  $s \in S_{ex} \Leftrightarrow$  когда посетитель  $v$  “ходил” на сайт  $s$ .

## Внутренняя таксономия

$\mathbb{K}_{in} = (V, S_{in}, I)$ , где

$V$  – множество всех посетителей целевого сайта,  $S_{in}$  – множество всех собственных страниц целевого сайта,  $I$  – отношение инцидентности  $vIs$ , имеющее место для  $v \in V$ ,  $s \in S_{in} \Leftrightarrow$  когда посетитель  $v$  “ходил” на сайт  $s$ .

- Понятие – пара  $(A, B)$
- $A' = \{ \text{множество сайтов } s \in S, \text{ которые посещали все посетители } v \in A \} = B$
- $B' = \{ \text{множество посетителей } v \in V, \text{ которые посещали все сайты } s \in B \} = A$ .

# Критерии отбора релевантных понятий

Пусть  $\mathbb{K} = (G, M, I)$  – формальный контекст,  $(A, B)$  – некоторое формальное понятие  $\mathbb{K}$ , тогда

## Индекс устойчивости

Индекс устойчивости  $\sigma$  понятия  $(A, B)$  определяется выражением

$$\sigma(A, B) = \frac{|\{C \subseteq A | C' = B\}|}{2^{|A|}}.$$

Очевидно, что  $0 \leq \sigma(A, B) \leq 1$ .

## Решетка-айсберг

Поддержка содержания понятия  $(A, B)$  определяется выражением  $supp(A, B) = \frac{|A|}{|G|}$ .

Пусть дано минимальное значение поддержки  $minsupp \in [0, 1]$ , тогда  
решеткой-айсбергом назовем множество  $\{(A, B) | supp(B) \geq minsupp\}$ .

# Исходные данные

- выборка по статистике посещений 10000 сайтов с прилагаемым плоским тематическим каталогом по 59 категориям.
- сайт университета, сайт Интернет-магазина бытовой техники, сайт крупного банка, сайт автомобильного Интернет-салона.

## Формат данных

id; \\id посетителя

first\_ts; \\время первого захода на сайт

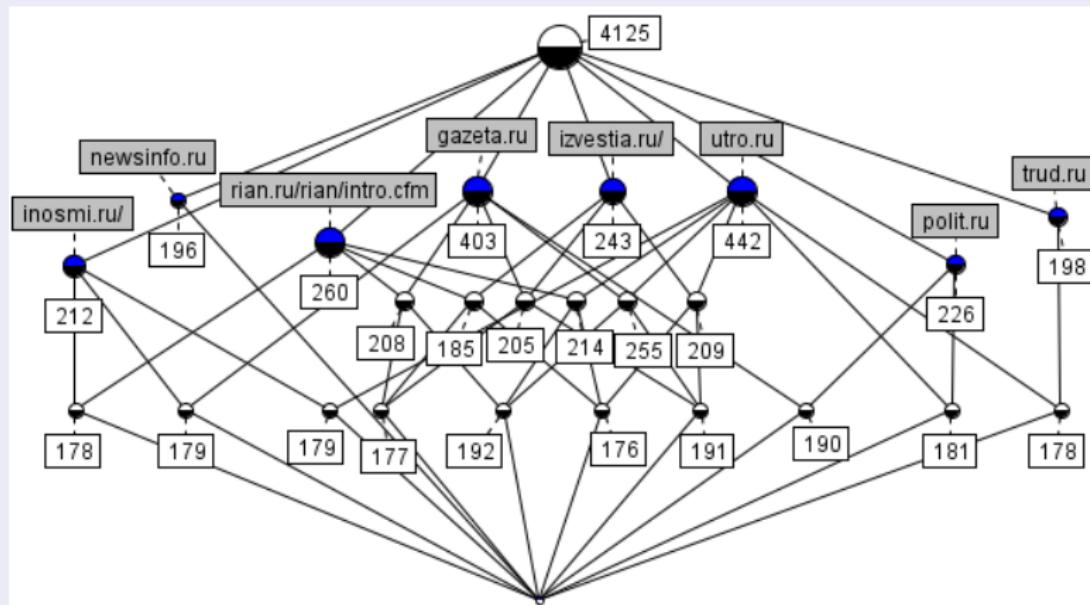
last\_ts; \\время последнего захода на сайт

num; \\количество совершенных сессий за все время знакомства с сайтом.

# Построение внешней таксономии

Сайт ВШЭ в сентябре 2006 года в терминах посещений новостных ресурсов.

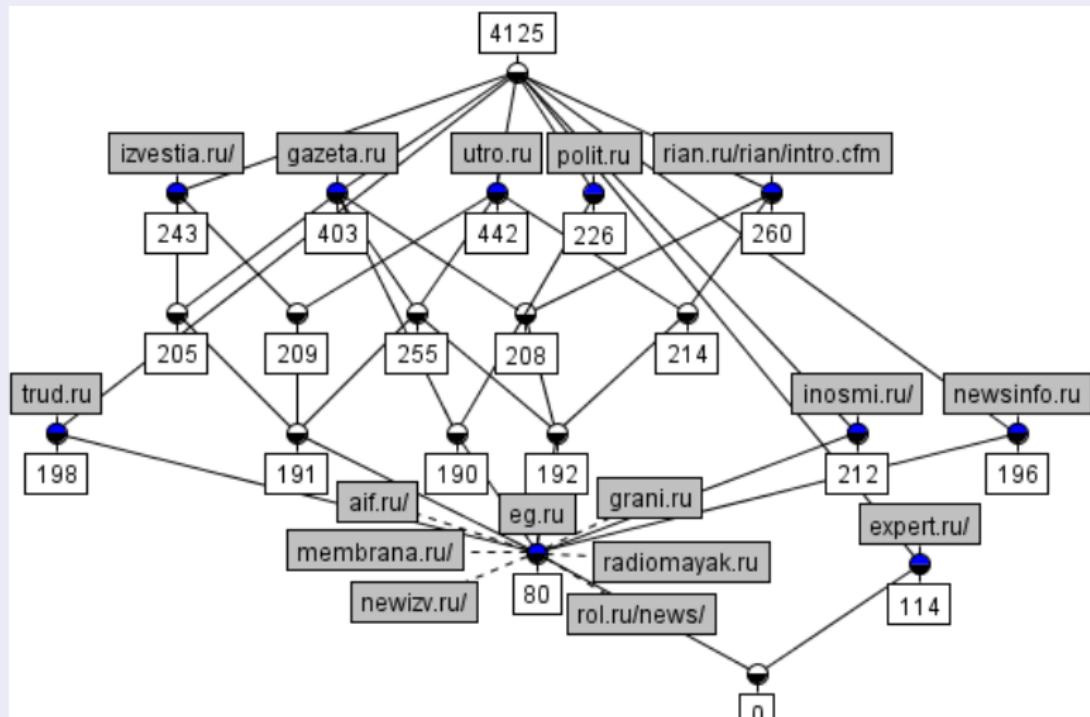
## Решетка-айсберг для 25 самых крупных понятий



# Построение внешней таксономии

Сайт ВШЭ в сентябре 2006 года в терминах посещений новостных ресурсов.

Диаграмма частично упорядоченного множества 25-и самых устойчивых понятий



# Рекомендация рекламных словосочетаний

- ① Разработка и реализация алгоритмов для формирования рекомендаций на массивах Интернет-данных
- ② Экспериментальная проверка применимости методов Data Mining для рекомендательной системы Интернет-рекламы

# Постановка задачи

- контекстная Интернет-реклама
- выявление рекламных слов, интересных рекламодателю
- пример — Google AdWords

# Рекомендация рекламных словосочетаний

## Исходные данные

Данные о покупках рекламных словосочетаний. Формальный контекст  
 $\mathbb{K}_{FT} = (F, T, I_{FT})$ ,  $F$  — множество компаний-рекламодателей,  $T$  — множество рекламных словосочетаний,  $fIt$  означает, что фирма  $f \in F$  купила словосочетание  $t \in T$ . Размер контекста —  $2000 \times 3000$ .

## Постановка задачи

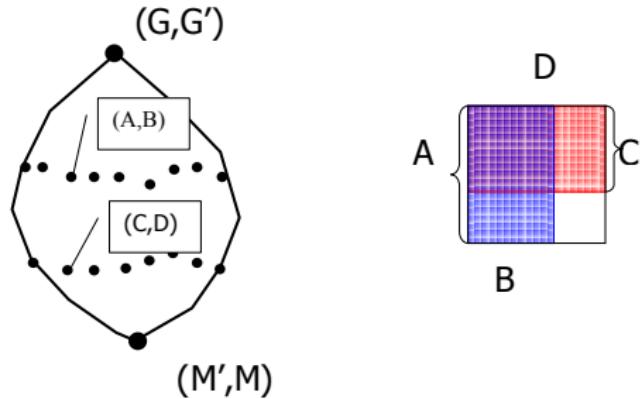
Требуется выявить рынки рекламных слов с целью последующего формирования рекомендаций

## Средства решения

- АФП: алгоритм D-miner
- поиск ассоциативных правил
- ассоциативные правила + морфология
- ассоциативные правила + онтология

# Рекомендация рекламных словосочетаний: АФП

[Besson et al, 2004], D-miner,  $O(|G|^2|M||L|)$



## Результаты работы алгоритма

Минимальный размер объема понятия	Минимальный размер содержания	Число формальных понятий
0	0	8 950 740
10	10	3 030 335
15	10	759 963
15	15	150 983
15	20	14 226
20	15	661

# Рекомендация рекламных словосочетаний: D-miner

## Рынок услуг по размещению сайтов

{affordable hosting web, business hosting web, cheap hosting, cheap hosting site web, cheap hosting web, company hosting web, cost hosting low web, discount hosting web, domain hosting, hosting internet, hosting page web, hosting service, hosting services web, hosting site web, hosting web}

## Гостиничный бизнес

{ angeles hotel los, atlanta hotel, baltimore hotel, dallas hotel, denver hotel, diego hotel san, francisco hotel san, hotel houston, hotel miami, hotel new orleans, hotel new york, hotel orlando, hotel philadelphia, hotel seattle, hotel vancouver }

# Рекомендация рекламных словосочетаний: ассоциативные правила

- [Szathmary, 2005]
- система Coron, алгоритм Zart, информативный базис ассоциативных правил

## Примеры правил

$\text{minsupp}=30$   $\text{minconf}=0,9$

- $\{\text{florist}\} \rightarrow \{\text{flower}\}$  supp=33 [1.65%]; conf=0.92;
- $\{\text{gift graduation}\} \rightarrow \{\text{anniversary gift}\}$ , supp=41 [2.05%]; conf=0.82;

## Результаты поиска ассоциаций

$\text{min\_supp}$	$\text{max\_supp}$	$\text{min\_conf}$	$\text{max\_conf}$	число правил
30	86	0,9	1	101 391
30	109	0,8	1	144 043

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

- $t$  — рекламное словосочетание,  $t = \{w_1, w_2, \dots, w_n\}$
- $s_i = stem(w_i)$  — основа слова  $w_i$
- $stem(t) = \bigcup_i stem(w_i)$  — множество основ словосочетания  $t$
- $\mathbb{K}_{TS} = (T, S, I_{TS})$  — формальный контекст, где  $T$  — множество всех словосочетаний, а  $S$  — множество основ всех словосочетаний из  $T$ , т.е.  $S = \bigcup_i stem(t_i)$
- $tIs$  означает, что во множество основ словосочетания  $t$  входит основа  $s$

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

Пример контекста  $\mathbb{K}_{FT}$  для рынка “long distance calling”

фирма \ фраза	call distance long	calling distance long	calling distance long plan	carrier distance long	cheap distance long
$f_1$	x		x		x
$f_2$		x	x	x	
$f_3$				x	x
$f_4$		x	x		x
$f_5$	x	x		x	x

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

Пример контекста  $K_{TS}$  для рынка “long distance calling”

фраза \ стем	call	carrier	cheap	distanc	long	plan
call distance long	x			x	x	
calling distance long	x			x	x	
calling distance long plan	x			x	x	x
carrier distance long		x		x	x	
cheap distance long			x	x	x	

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

## Примеры

- $t \xrightarrow{FT} s_i^{ITS}$

$\{last\ minute\ vacation\} \rightarrow \{last\ minute\ travel\}$

Supp= 19 Conf= 0,90

- $t \xrightarrow{FT} \bigcup_i s_i^{ITS}$

•  $\{mail\ order\ phentermine\} \rightarrow$

$\{adipex\ online\ order, adipex\ order, adipex\ phentermine, \dots,$   
 $phentermine\ prescription, phentermine\ purchase, phentermine\ sale\}$

Supp= 19 Conf= 0,95

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

## Примеры

- $t \xrightarrow{FT} (\bigcup_i s_i)^{I_{TS}}$
- $\{distance\ long\ phone\} \rightarrow$   
 $\{call\ distance\ long\ phone, carrier\ distance\ long\ phone, \dots,$   
 $distance\ long\ phone\ rate, distance\ long\ phone\ service\}$   
Supp= 37 Conf= 0,88
- $t_1 \xrightarrow{FT} t_2$ , такие что  $t_2^{I_{TS}} \subseteq t_1^{I_{TS}}$
- $\{ink\ jet\} \rightarrow \{ink\}$ , Supp= 14 Conf= 0,7

# Рекомендация рекламных словосочетаний: ассоциативные правила+морфология

$min\_conf = 0.5$

## Проверка качества правил

Тип правила	Среднее значение supp	Среднее значение conf	Число правил
$t \xrightarrow{FT} s_i^{I_{TS}}$	15	0,64	454
$t \xrightarrow{FT} \bigcup_i s_i^{I_{TS}}$	15	0,63	75
$t \xrightarrow{FT} (\bigcup_i s_i)^{I_{TS}}$	18	0,67	393
$t \xrightarrow{FT} t_i, \text{ где } t_i^{I_{TS}} \subseteq t^{I_{TS}}$	21	0,70	3922
$t \xrightarrow{FT} \bigcup_i t_i, \text{ где } t_i^{I_{TS}} \subseteq t^{I_{TS}}$	20	0,69	673

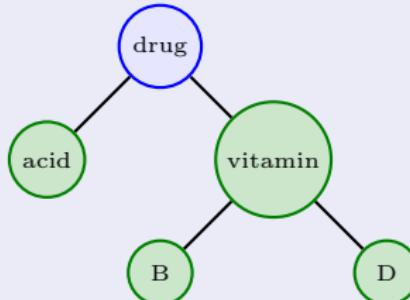
# Рекомендация рекламных словосочетаний: ассоциативные правила

## Результаты скользящего контроля для ассоциативных правил

	Число правил	Число правил с $\text{sup} > 0$	average_conf	Число правил с $\text{min\_conf}=0.5$	average_conf ( $\text{min\_conf}=0.5$ )
1	147170	73025	0,77	65556	0,84
2	69028	68709	0,93	68495	0,93
3	89332	89245	0,95	88952	0,95
4	107036	93078	0,84	86144	0,90
5	152455	126275	0,82	113008	0,90
6	117174	114314	0,89	111739	0,91
7	131590	129826	0,95	128951	0,96
8	134728	120987	0,96	106155	0,97
9	101346	67873	0,72	52715	0,92
10	108994	107790	0,93	106155	0,94
средние	115885	99112	0,87	92787	0,92

# Рекомендация рекламных словосочетаний: ассоциативные правила+онтология

## Составление онтологии (иерархического каталога)



## Метаправила и соответствующие им ассоциации

- сопоставление правилам онтологии ассоциаций
- $t \rightarrow g_i(t)$ , где  $g_i(t)$  — множество понятий онтологии на  $i$  уровне выше  $t$
- $t \rightarrow n(t)$ , где  $n(t)$  — множество соседних для  $t$  понятий онтологии, имеющих общего предка

# Рекомендация рекламных словосочетаний: ассоциативные правила+онтология

## Примеры правил

- $t \rightarrow g_1(t)$
- $\{d\text{ vitamin}\} \rightarrow \{\text{vitamin}\}$ , Supp= 19 Conf= 0,90
- $t \rightarrow n(t)$
- $\{b\text{ vitamin}\} \rightarrow \{ b\text{ complex vitamin}, b12\text{ vitamin}, c\text{ vitamin}, d\text{ vitamin}, discount\text{ vitamin}, e\text{ vitamin}, herb\text{ vitamin}, mineral\text{ vitamin}, multi\text{ vitamin}, supplement\text{ vitamin}\}$  Supp= 18 Conf= 0,7

# Содержание

## 1 Основная часть

- Введение
- Области применения
- Анализ формальных понятий
- Частые множества и ассоциативные правила
- Алгоритм Apriori
- Алгоритм FP-growth
- Меры интересности правил
- Компактное представление частых множеств

## 2 Прикладные задачи и эксперименты

- Анализ посещаемости веб-сайтов
- Рекомендация контекстной рекламы

## 3 Программные средства

## 4 Чего бы почитать и посмотреть?

# Основные свободно-распространяемые инструменты

- SPMF – an open-source data mining mining library
- The CORON Data Mining Platform
- Bart Goethals webpage and FIMI repository
- Conexp — решетки понятий, импликации и ассоциативные правила
- Orange – содержит виджеты для поиска частых множеств признаков и ассоциативных правил (версия 2.7)
- Spark ML Lib – frequent itemset mining via FP-growth and association rules
- Frequent Itemset Mining in Python

# Содержание

## 1 Основная часть

- Введение
- Области применения
- Анализ формальных понятий
- Частые множества и ассоциативные правила
- Алгоритм Apriori
- Алгоритм FP-growth
- Меры интересности правил
- Компактное представление частых множеств

## 2 Прикладные задачи и эксперименты

- Анализ посещаемости веб-сайтов
- Рекомендация контекстной рекламы

## 3 Программные средства

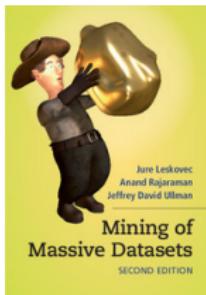
## 4 Чего бы почитать и посмотреть?

# Книги

- M. Zaki et al. [Data Mining and Analysis: Fundamental Concepts and Algorithms](#), 2014 (free)
- J. Leskovec et al. [Mining of Massive Datasets](#), 2014 (free)
- J. Han et al. [Data Mining. Concepts and Techniques](#), 2012
- Барсегян А. и др. [Анализ данных и процессов](#), 2009

# Coursera: курсы и специализации

<http://www.coursera.org/>



- Jiawei Han Pattern Discovery in Data Mining (current)
- Jure Leskovec et al. Mining Massive Datasets (current)

Специализации (платные сертификаты) — состоят из отдельных курсов (участие бесплатно)

- Data Mining (current)

- Интернет-университет информационных технологий
- К.В. Воронцов [Машинное обучение](#), 2015 (Видео к курсу на сайте ШАД)
- И.А. Чубукова. [Data Mining](#), 2006

# Сообщество и источники данных

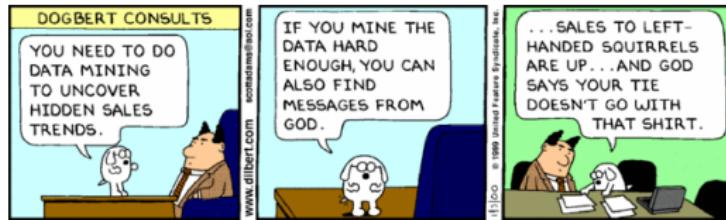
- IMLS – The International Machine Learning Society
- Kaggle – платформа для соревнований по анализу данных
- KDD Nuggets – Data Mining Community Top Resource
- Open ML – Machine Learning community portal
- UCI Machine Learning Repository – Репозиторий данных

# Конференции

- IEEE ICDM – IEEE International Conference on Data Mining
- KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- ECML & PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- AICT – International conference on Analysis of Images, Social Networks, and Texts

# Just for fun или шутки ради

<http://dilbert.com>



# Вопросы и контакты

[www.hse.ru/staff/dima](http://www.hse.ru/staff/dima)

Спасибо!

dmitrii.ignatov[at]gmail.com