

Решетки формальных понятий в современных методах анализа данных и знаний

С.О.Кузнецов

Национальный исследовательский университет высшая школа
экономики
Москва
Большие данные в национальной экономике
Москва

22 октября 2013 г.

Поиск знаний в больших сложных и неточных данных

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Растущая потребность в анализе больших, сложных и неточных данных в биоинформатике, анализе текстов и изображений
- Неточность в данных может естественным образом задаваться интервалами значений
- Анализ формальных понятий и его расширение - узорные структуры (pattern structures) дают средства для анализа такого рода данных

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

АФП в задачах поиска знаний

Ассоциативные правила в майнинге данных

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Алгоритм Apriori [R.Agrawal et al. 1993] поиска частых ассоциативных правил вида $A \rightarrow_{\text{conf, supp}} B$
- Более эффективные алгоритмы на основе замкнутых множеств признаков [L.Lakhal et al. 2000], [L.Szathmary et al. 2007], etc.
- Распространение идеи использования замкнутых множеств признаков на произвольные замкнутые описания

Basic notions of Formal Concept Analysis

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

[Wille 1982], [Ganter, Wille 1996]

- M , a set of **attributes**
- G , a set of **objects**
- relation $I \subseteq G \times M$ such that $(g, m) \in I$ if and only if the object g has the attribute m .
- $\mathbb{K} := (G, M, I)$ is a **formal context**.

Derivation operators: $A \subseteq G, B \subseteq M$

$$A' \stackrel{\text{def}}{=} \{m \in M \mid glm \text{ for all } g \in A\}, \quad B' \stackrel{\text{def}}{=} \{g \in G \mid glm \text{ for all } m \in B\}$$

A **formal concept** is a pair (A, B) : $A \subseteq G, B \subseteq M, A' = B$, and $B' = A$.

- A is the **extent** and B is the **intent** of the concept (A, B) .
- The concepts, ordered by $(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2$ form a complete lattice, called **the concept lattice** $\mathfrak{B}(G, M, I)$.

Example. Diagram of the ordered set of concepts

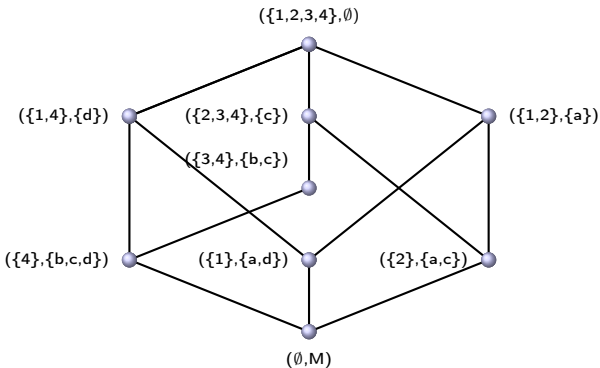
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний





С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



	G \ M	a	b	c	d
1		×			×
2		×		×	
3			×	×	
4			×	×	×

- a** – has 3 vertices,
- b** – has 4 vertices,
- c** – has a direct angle,
- d** – equilateral

Implications

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- **Implication** $A \rightarrow B$ for $A, B \subseteq M$ holds if $A' \subseteq B'$, i.e., every object that has all attributes from A also has all attributes from B .
- **Armstrong rules:**

$$\frac{A \rightarrow B}{A \cup C \rightarrow B} \quad , \quad \frac{A \rightarrow B, A \rightarrow C}{A \rightarrow B \cup C} \quad , \quad \frac{A \rightarrow B, B \rightarrow C}{A \rightarrow C}$$

- **A Minimal implication base:**
A base with the minimum number of implications [Duquenne, Guigues 1986] or
the **stem base**, its premises can be given (Ganter 1987) by pseudointents:
 - A set $P \subseteq M$ is a **pseudointent** if
$$P \neq P'' \text{ and } Q'' \subset P \text{ for every pseudointent } Q \subset P.$$

Aliases of Implication

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Functional dependency
- Jumping emergent pattern (for implications with a fixed right-hand side)
- Exact association rule
- Element of a (disjunctive) version space (for implications with a fixed right-hand side) [T.Mitchell 1982] , M.Sebag [1993]
- Elements of Horn theories [Angluin et al. 1992], [Kautz et al. 1993]

Association rules

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Partial implications [M. Luxenburger, 1991]

Association rules [Agrawal et al., 1993]

A partial implication (association rule) of context (G, M, I) is an expression $A \rightarrow_{c,s} B$, where

- $c, s \in [0, 1]$;
- $c = \frac{|(A \cup B)'|}{|A'|}$, called **confidence**, $\text{conf}(A \rightarrow B)$;
- $s = \frac{|(A \cup B)'|}{|G|}$, called **support**, $\text{supp}(A \rightarrow B)$.

Implication is an association rule with confidence = 1.

Concise representation of association rules [L.Lakhal et al. 2000]

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Find all "frequent" (with support greater than a threshold) association rules with confidence greater than a threshold.

Solution stages

- Find all frequent "closed itemsets"(frequent intents)
- For each frequent intent B find all its maximal subintents A_1, \dots, A_n
- Retain only those A_i for which $\text{conf}(A_i \rightarrow B) \geq \theta$, where θ is confidence threshold
- Find minimal generators of the remaining A_i , compose rules of the form $\text{mingen}(A_i) \rightarrow B$.

Luxenburger basis

- Spanning tree of the concept lattice diagram
- Duquenne-Guigues implication base

Example. Diagram of the ordered set of concepts.

Confidence

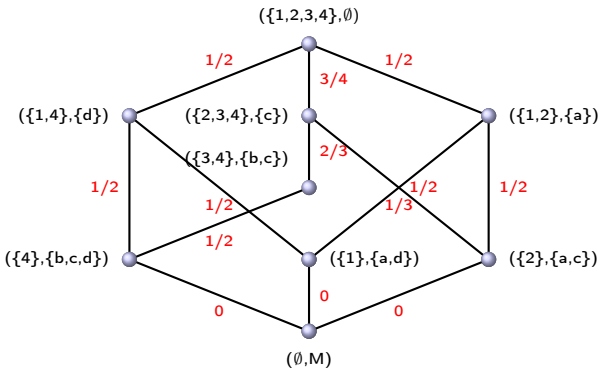
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний





С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



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

Good rules with $supp \geq 1/2$ and $minconf \geq 3/4$

1. $\emptyset \rightarrow c$, $sup(\emptyset \rightarrow c) = conf(\emptyset \rightarrow c) = 3/4$;

2. $c \rightarrow b$, $sup(c \rightarrow b) = 1/2$, $conf(c \rightarrow b) = 2/3$.

Example. Diagram of the ordered set of concepts.

Support

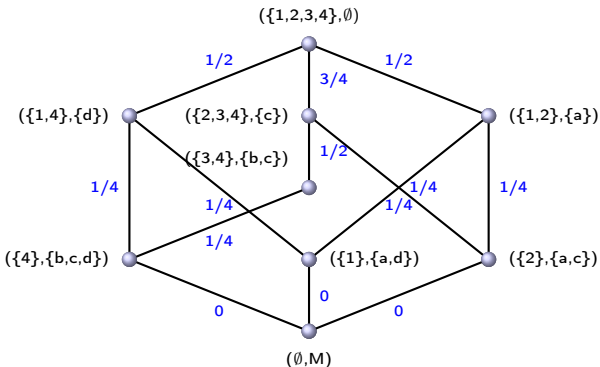
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



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

Good rules with $supp \geq 1/2$ and $minconf \geq 3/4$

- $\emptyset \rightarrow c$, $sup(\emptyset \rightarrow c) = conf(\emptyset \rightarrow c) = 3/4$;
- $c \rightarrow b$, $sup(c \rightarrow b) = 1/2$, $conf(c \rightarrow b) = 2/3$.

JSM-hypotheses

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

[Finn 1991], [Ganter, Kuznetsov 2000]

A **target attribute** $w \notin M$,

- **positive examples:** Set $G_+ \subseteq G$ of objects known to have w ,
- **negative examples:** Set $G_- \subseteq G$ of objects known not to have w ,
- **undetermined examples:** Set $G_\tau \subseteq G$ of objects for which it is unknown whether they have the target attribute or do not have it.

Three subcontexts of $\mathbb{K} = (G, M, I)$: $\mathbb{K}_\varepsilon := (G_\varepsilon, M, I_\varepsilon)$, $\varepsilon \in \{-, +, \tau\}$ with respective derivation operators $(\cdot)^+$, $(\cdot)^-$, and $(\cdot)^\tau$.

A **positive hypothesis** $H \subseteq M$ is an intent of \mathbb{K}_+ not contained in the intent g^- of any negative example $g \in G_-$: $\forall g \in G_- \quad H \not\subseteq g^-$. Equivalently,

$$H^{++} = H, \quad H' \subseteq G_+ \cup G_\tau.$$

Example of a learning context

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

	G \ M	color	firm	smooth	form	fruit
1	apple	yellow	no	yes	round	+
2	grapefruit	yellow	no	no	round	+
3	kiwi	green	no	no	oval	+
4	plum	blue	no	yes	oval	+
5	toy cube	green	yes	yes	cubic	-
6	egg	white	yes	yes	oval	-
7	tennis ball	white	no	no	round	-
8	mango	green	no	yes	oval	τ

Natural scaling of the context

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

	G \ M	w	y	g	b	f	\bar{f}	s	\bar{s}	r	\bar{r}	fruit
1	apple		x				x	x		x		+
2	grapefruit		x				x		x	x		+
3	kiwi			x			x		x		x	+
4	plum				x		x	x			x	+
5	toy cube			x		x		x			x	-
6	egg	x				x		x			x	-
7	tennis ball	x					x		x	x		-
8	mango			x			x	x			x	τ

Abbreviations:

"g" for green, "y" for yellow, "w" for white, "f" for firm, " \bar{f} " for nonfirm,
 "s" for smooth, " \bar{s} " for nonsmooth, "r" for round,
 " \bar{r} " for nonround.

Positive Concept Lattice

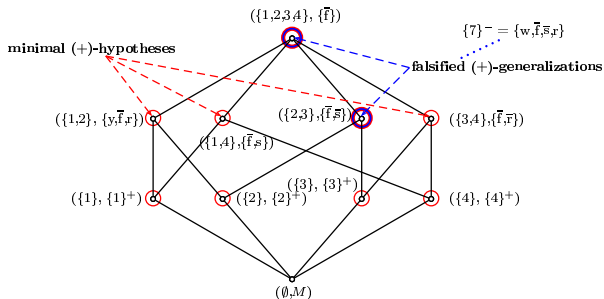
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



	G \ M	w	y	g	b	f	f̄	s	s̄	r	r̄	fruit
1	apple		x				x	x		x		+
2	grapefruit		x				x		x	x		+
3	kiwi			x			x		x		x	+
4	plum				x		x	x			x	+
5	toy cube			x		x		x			x	-
6	egg	x				x		x			x	-
7	tennis ball	x					x		x	x		-
8	mango			x			x	x			x	τ

Hypotheses vs. implications

A positive hypothesis h corresponds to an implication $h \rightarrow \{w\}$ in the context $K_+ = (G_+, M \cup \{w\}, I_+ \cup G_+ \times \{w\})$.

A negative hypothesis h corresponds to an implication $h \rightarrow \{\bar{w}\}$ in the context $K_- = (G_-, M \cup \{\bar{w}\}, I_- \cup G_- \times \{\bar{w}\})$.

Hypotheses are special implications: their premises are closed (in K_+ or in K_-).

	G \ M	w	y	g	b	f	\bar{f}	s	\bar{s}	r	\bar{r}	fruit
1	apple		x				x	x		x		x
2	grapefruit		x				x		x	x		x
3	kiwi			x			x		x		x	x
4	plum				x		x	x			x	x
5	toy cube			x		x		x			x	x
6	egg	x				x		x			x	x
7	tennis ball	x					x		x	x		x

Stability index for selecting best concepts

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Definition

Given a concept (A, B) , (intentional) concept stability $Stab(A, B)$ is the fraction of subextents of A that have the same intent B :

$$Stab(A, B) := \frac{|\{D \subseteq A \mid D' = B\}|}{2^{|A|}}$$

Example: Two contexts with three objects each.

$$g'_1 = \{b, n, c\} \quad g'_2 = \{b, n, e\} \quad g'_3 = \{b, f\}$$

$$Stab(\{1, 2, 3\}, \{b\}) = 3/8$$

$$g'_4 = \{b, n, c\} \quad g'_5 = \{b, e\} \quad g'_6 = \{b, f\}$$

$$Stab(\{4, 5, 6\}, \{b\}) = 4/8$$

Toxicology analysis by means of hypotheses

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

V.G.Blinova, D.A.Dobrynin, V.K.Finn, S.O.Kuznetsov, E.S.Pankratova,
Toxicology Analysis by Means of JSM-Method, *Bioinformatics*, 19(2003)

Training Set: Data of the National Toxicology Program (NTP) with 120 to 150 positive examples and 190 to 230 negative examples of toxicity: molecular graphs with indication of whether a substance is toxic for four sex/species groups: {male, female} \times {mice, rats}.

Test Set: Data of Food and Drug Administration (FDA): about 200 chemical compounds with known molecular structures, whose (non)toxicity, known to organizers, was to be predicted by participants.

Participants: 12 research groups (world-wide), each with up to 4 prediction models for every sex/species group.

Evaluation: ROC diagrams

Stages of the Competition:

1. Encoding of chemical structures in terms of attributes,
2. Generation of classification rules,
3. Prediction by means of classification rules.

Results of each stage were made public (put on a web site).

Fragmentary Code of Substructure Superposition

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

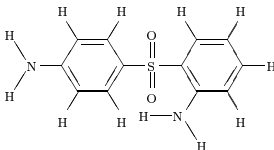
Узорные
структуры и
вызов
больших
данных

Example of Coding with FCSS

[Avidon et al. 1982], [Blinova et al., 2000]

Chemical structure

Complete list of descriptors

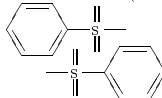


6,06	×2
0200331	×2
1300241	×2
2400331	×2
0264241	
0262241	

6,06 (cyclic descriptors)




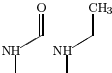
0200331 (linear descriptors)



ILP'05 [23]

Positive hypotheses for toxicity

Some positive hypotheses for toxicity causes

Molecular graph	FCCS descriptors (encoding)	# of predictions in sex/species group(s)
	{(6,06), (0200021)}	2 for FR
	{(0201131), (0202410)}	1 for FR, 1 for MM

Roc diagrams: Female mice

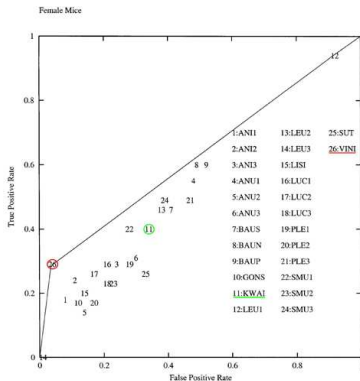
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Roc diagrams: Male mice

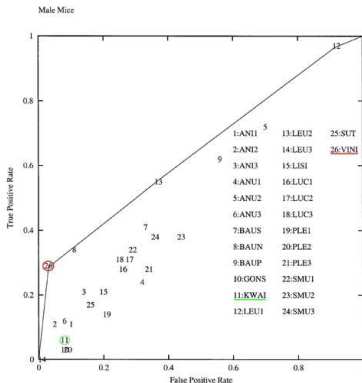
Решетки
формаль-
ных поня-
тий в со-
времен-
ных мето-
дах ана-
лиза
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Roc diagrams: Female rats

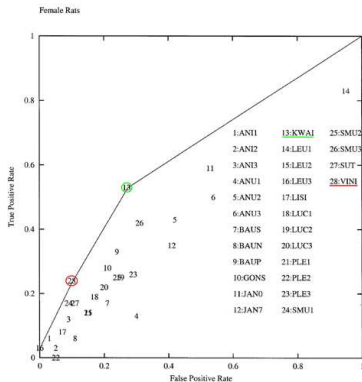
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Roc diagrams: Male rats

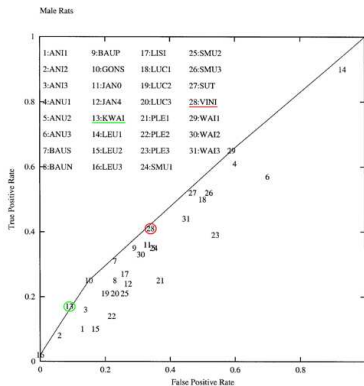
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Узорные структуры

Beyond datatables: graphs with labels

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

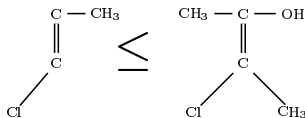
Узорные
структуры и
вызов
больших
данных

Let (\mathcal{L}, \preceq) be an ordered set of vertex labels.

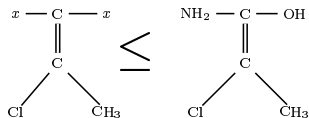
$\Gamma_1 := ((V_1, l_1), E_1)$ **dominates** $\Gamma_2 := ((V_2, l_2), E_2)$ or $\Gamma_2 \leq \Gamma_1$
if there exists a one-to-one mapping $\varphi: V_2 \rightarrow V_1$ such that

- respects edges: $(v, w) \in E_2 \Rightarrow (\varphi(v), \varphi(w)) \in E_1$,
- fits under labels: $l_2(v) \preceq l_1(\varphi(v))$.

Example: $\mathcal{L} = \{x, NH_2, Cl, CH_3, C, OH\}$



vertex labels are unordered



$x \preceq A$ for any vertex label $A \in \mathcal{L}$

Semilattice on graph sets

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

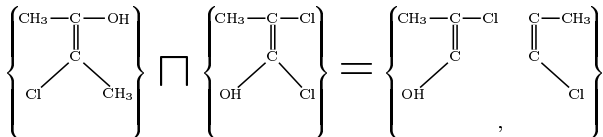
Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

$$\{X\} \sqcap \{Y\} := \{Z \mid Z \leq X, Y, \quad \forall Z_* \leq X, Y \quad Z_* \not\leq Z\}$$

= The set of all maximal common subgraphs of X and Y .

Example:



Meet of graph sets

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

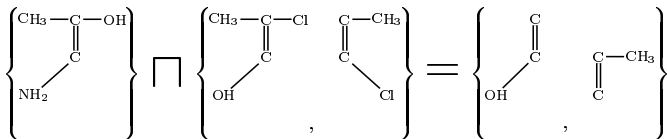
For sets of graphs

$$\mathcal{X} = \{X_1, \dots, X_k\} \text{ and } \mathcal{Y} = \{Y_1, \dots, Y_n\}$$

$$\mathcal{X} \sqcap \mathcal{Y} := \text{MAX}_{\leq}(\cup_{i,j}(\{X_i\} \sqcap \{Y_j\}))$$

\sqcap is idempotent, commutative, and associative.

Example:



Pattern Structure

[Ganter, Kuznetsov 2001]

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

$(G, \underline{D}, \delta)$ is a **pattern structure** if

- G is a set ("set of objects");
- $\underline{D} = (D, \sqcap)$ is a meet-semilattice;
- $\delta : G \rightarrow D$ is a mapping;
- the set $\delta(G) := \{\delta(g) \mid g \in G\}$ generates a complete subsemilattice (D_δ, \sqcap) of (D, \sqcap) .

Possible origin of \sqcap operation:

- A set of objects G , each with description from P ;
- Partially ordered set (P, \leq) of "descriptions" (\leq is a "more general than" relation);
- The (distributive) lattice of order ideals of the ordered set (P, \leq) .

Pattern Structure

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Let $(G, (D, \sqcap), \delta)$ be a **pattern structure**, then

the subsumption order is defined as $c \sqsubseteq d: = c \sqcap d = c$.

Derivation operators:

$$A^\diamond := \sqcap_{g \in A} \delta(g) \text{ for } A \subseteq G$$

$$c^\diamond := \{g \in G \mid c \sqsubseteq \delta(g)\} \text{ for } c \in C.$$

A pair (A, c) is a **pattern concept** of $(G, (C, \sqcap), \delta)$ if

$$A \subseteq G, c \in C, A^\diamond = c, c^\diamond = A$$

A is **extent** and c is **pattern intent**.

$A \subseteq G$ is **closed** if $A^{\diamond\diamond} = A$.

$d \in D$ is **closed** if $d^{\diamond\diamond} = d$.

Implications and Associations in Pattern Structures

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

For $A, B \in D$

Implication $A \rightarrow B$ holds if $A^\diamond \subseteq B^\diamond$.

Association rule $A \rightarrow B$ with support s and confidence c holds if

$$c = \frac{|(A^\diamond \cap B^\diamond)|}{|A^\diamond|}$$

$$s = \frac{|(A^\diamond \cap B^\diamond)|}{|G|}.$$

A training sample

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

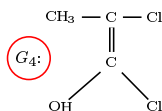
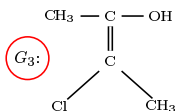
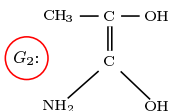
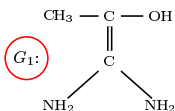
С.О.Кузнецов

АФП для
поиска
знаний в
данных

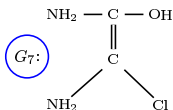
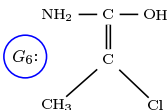
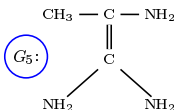
Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Positive examples:



Negative examples:



Positive lattice

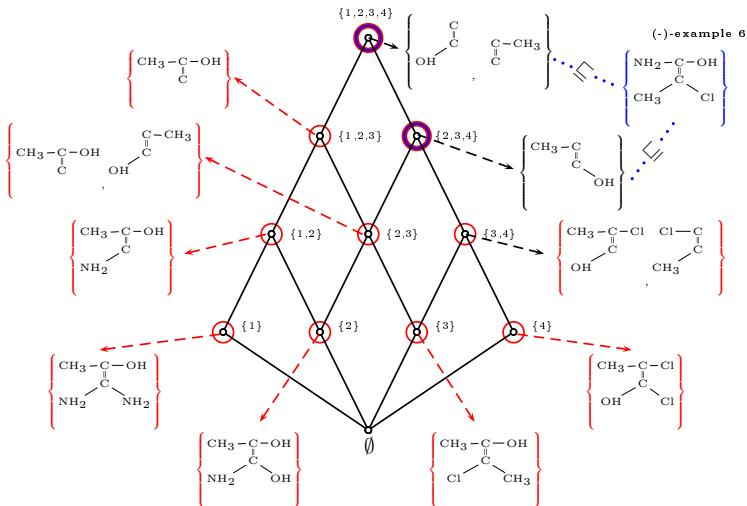
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Pattern Structures

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Let $(G, (D, \sqcap), \delta)$ be a **pattern structure**, then

the subsumption order is defined as $c \sqsubseteq d: = c \sqcap d = c$.

Derivation operators:

$$A^\diamond := \sqcap_{g \in A} \delta(g) \text{ for } A \subseteq G$$

$$c^\diamond := \{g \in G \mid c \sqsubseteq \delta(g)\} \text{ for } c \in C.$$

A pair (A, c) is a **pattern concept** of $(G, (C, \sqcap), \delta)$ if

$$A \subseteq G, c \in C, A^\diamond = c, c^\diamond = A$$

A is **extent** and c is **pattern intent**.

$A \subseteq G$ is **closed** if $A^{\diamond\diamond} = A$.

$d \in D$ is **closed** if $d^{\diamond\diamond} = d$.

Reinventing the closure in Data Mining

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Late 1990s: a wave of interest in graph mining, with application in chemistry, protein analysis, analysis of XML documents, etc. First Apriori-like algorithm gSpan was fairly efficient.

X. Yan and J. Han, gSpan: Graph-Based Substructure Pattern Mining, Proc. IEEE Int. Conf. on Data Mining, ICDM'02, 2002, pp.721–724, IEEE Computer Society

However, it was outperformed by CloseGraph from

X. Yan and J. Han, CloseGraph: mining closed frequent graph patterns, *Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'03*

where **closed graphs** are defined in terms of “counting inference”:

Given a labeled graph dataset D and a graph $g \in D$

support(g) is a set (or number) of graphs in D , in which g is a subgraph.

A graph g is called **closed** if no supergraph f of g (i.e., a graph such that g is isomorphic to its subgraph) has the same support.

Closed graphs and closed sets of graphs

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Remark: Unlike closed subsets of attributes, closed graphs, ordered by subgraph isomorphism, do not make a lattice (there can be multiple sups and infs).

However, closed graphs are related to closed sets of graphs (i.e., sets \mathcal{G} such that $\mathcal{G}^{\diamond\diamond} = \mathcal{G}$) as follows:

Proposition. Let a labeled graph dataset D be given, then

1. For a closed graph g there is a closed set of graphs \mathcal{G} such that $g \in \mathcal{G}$.
2. For a closed set of graphs \mathcal{G} and an arbitrary $g \in \mathcal{G}$, graph g is closed.

Projections as Approximation Tool

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Motivation: Complexity of computations in $(G, \underline{D}, \delta)$, e.g., testing SUBGRAPH ISOMORPHISM, i.e., relation \leq for graphs, is NP-complete.

ψ is **projection** on an ordered set (D, \sqsubseteq) if ψ is

monotone: if $x \sqsubseteq y$, then $\psi(x) \sqsubseteq \psi(y)$,

contractive:

a. $\psi(x) \sqsubseteq x$

b. $\forall x, \forall y, \exists z: y \sqsubseteq \psi(x) \Rightarrow y = \psi(z)$

idempotent: $\psi(\psi(x)) = \psi(x)$.

Projections as an Approximation Tool

Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

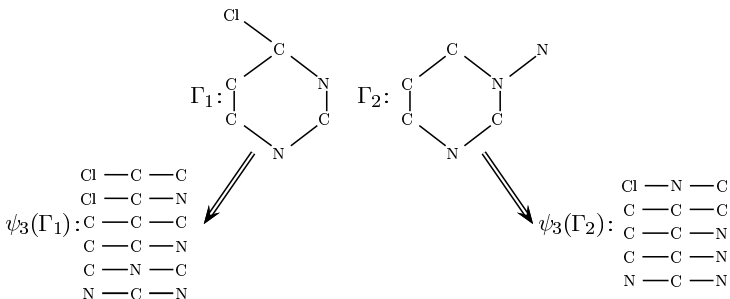
С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Example. Projection $\psi_3(\Gamma)$ takes Γ_1 and Γ_2 to the sets of their connected 3-vertex subgraphs.

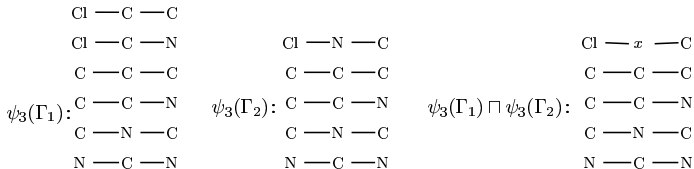


A property of projections

Any projection of a complete semilattice (D, \sqcap) is \sqcap -preserving, i.e., for any $X, Y \in D$

$$\psi(X \sqcap Y) = \psi(X) \sqcap \psi(Y).$$

Example. A projection $\psi_n(\Gamma)$ takes Γ to the set of its n -chains not dominated by other n -chains. Here $n = 3$, the label x is smaller than other labels, other labels are pairwise incomparable.



Projections and Representation Context

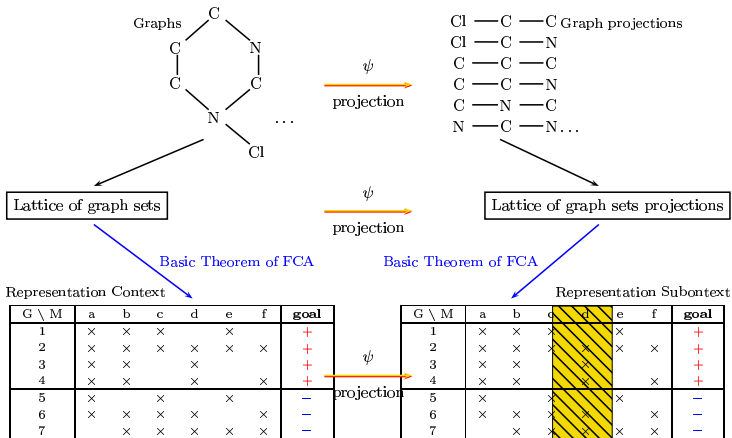
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Projection types used in chemical applications

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

We used several types of projections of labeled graph sets that are natural in chemical applications:

- *k-chain* projection: a set of graphs X is taken to the set of all chains with k vertices that are subgraphs of at least one graph of the set X ;
- *k-vertex* projection: a set of graphs X is taken to the set of all subgraphs with k vertices that are subgraphs of at least one graph of the set X ;
- *k-cycles* projection: a set of graphs X is taken to the set of all subgraphs consisting of k adjacent cycles of a minimal cyclic basis of at least one graph of the set X .

Mixed projections (with same algebraical properties of simple projections) are also possible.

4-Projections

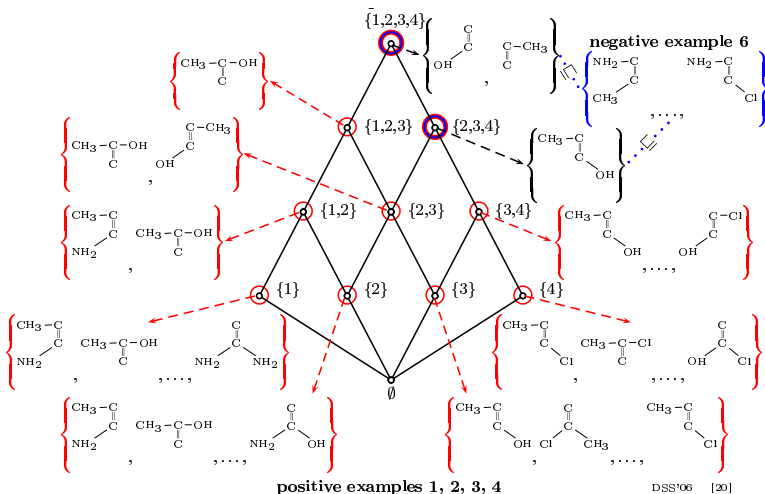
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



3-Projections

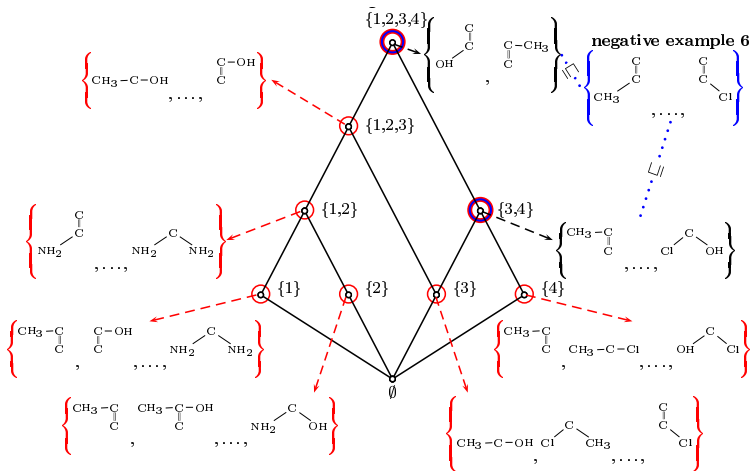
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



positive examples 1, 2, 3, 4

DSS'06 [21]

2-Projections

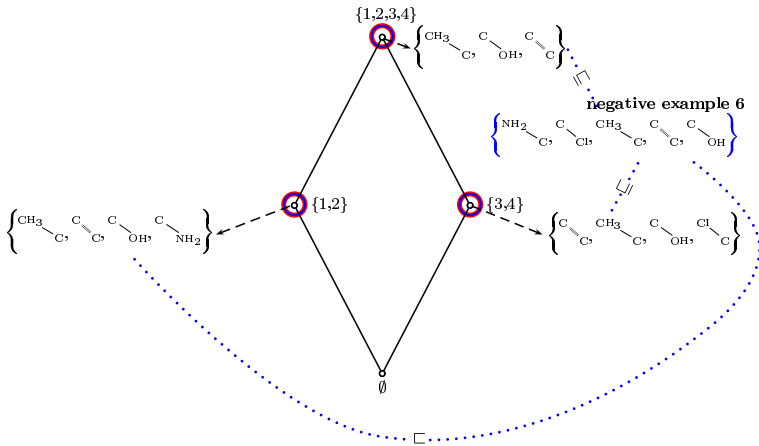
Решетки
формальных
понятий в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



positive examples 1, 2, 3, 4

DSS'06 [22]

Improving results with graph patterns: Predicting toxicity for female rats

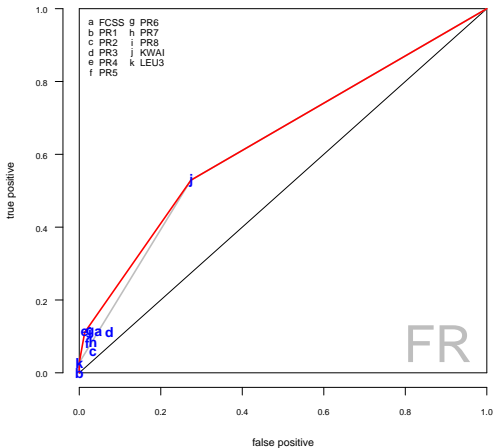
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Improving results with graph patterns: Predicting toxicity for male rats

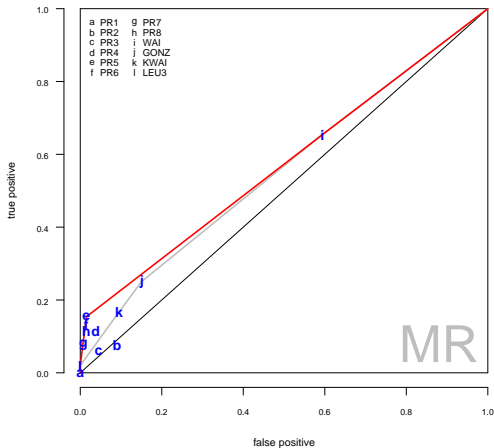
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Applications of graph patterns in chemoinformatics

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- 1 Predicting biological properties of chemical compounds (toxicity of alcohols, carcinogenicity of halogen substituted hydrocarbons and polycyclic aromatic hydrocarbons, etc.)
- 2 Forecast of biotransformation pathways in human organism
- 3 Analysis of chemical reactions

Other applications of pattern structures

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- 1 String patterns for studying protein and DNA sequences first in [S.Ferré et al., 2003] within Logical Concept Analysis
- 2 String patterns for (business) process mining [A.Buzmakov et al. 2012]
- 3 Interval vector patterns for analyzing gene expression data [M.Kaytoue et al. 2009]
- 4 Text clustering, classification, and retrieval based on graph representation [B.Galitsky et al. 2013]

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Узорные структуры на графах для анализа и майнинга текстов

Similarity between two paragraphs of text

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Baseline: bag-of-words approach, which computes the set of common keywords/n-grams and their frequencies.
- Pair-wise matching of syntactic parse trees representing individual sentences.
- Paragraph-paragraph match retaining the structure of the paragraphs

Finding similarity between two paragraphs

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

"Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons",

"UN nuclear watchdog passes a resolution condemning Iran for developing a second uranium enrichment site in secret",

"A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons",

"Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US"

□

"UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claims that its nuclear research is for peaceful purpose",

"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret",

"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"

Keywords: topic with no details

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Iran, UN, proposal, dispute, nuclear, weapons, passes,
resolution, developing, enrichment, site, secret, condemning,
second, uranium

Improvement: pair-wise generalization of syntactic trees

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

[NN-work IN-* IN-on JJ-nuclear NNS-weapons], [**DT-the NN-dispute
IN-over JJ-nuclear NNS-***], [VBZ-passes DT-a NN-resolution],
[VBG-condemning NNP-iran IN-*],
[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]],
[DT-* JJ-second NN-uranium NN-enrichment NN-site]],
[VBZ-is IN-for JJ-peaceful NN-purpose],
[DT-the NN-evidence IN-* PRP-it], [**VBN-* VBN-fabricated IN-by
DT-the NNP-us**]

Parse Thickets

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Parse thicket of a paragraph [B.Galitsky et al. 2013] is the set of parse trees of the sentences comprising the paragraph augmented by links between words from different sentences. These links may originate from

- Anaphora
- Same entity
- Hyponym/Hyperonym relation
- Rhetoric Structure Theory (RST) [Mann]
- Speech Act Theory (communicative actions, CA) [Searle]

A natural useful projection of a parse thicket is the set of its phrases (noun phrases, verb phrases, etc.)

Parse Thicket. An example

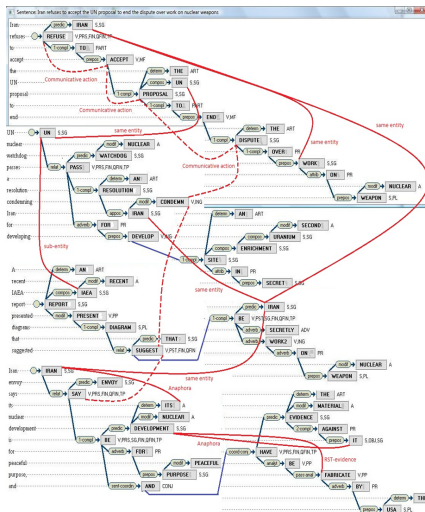
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Similarity of parse thickets. An example

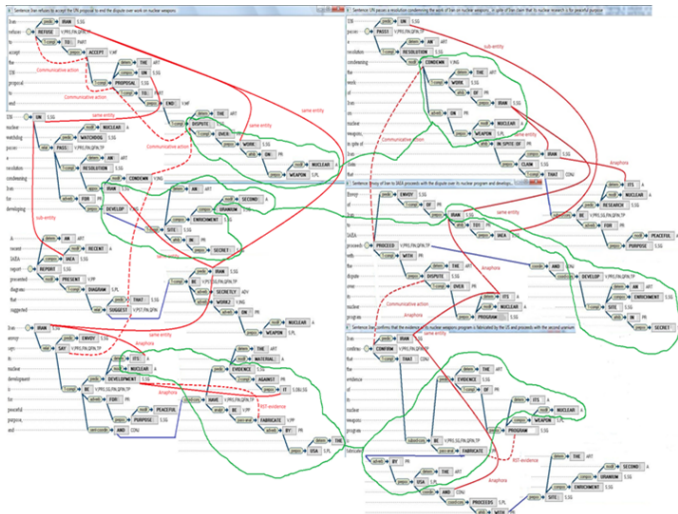
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Three domains were used in evaluation [B.Galitsky et al. 2013]:

- Product recommendation, where an agent reads chats about products and finds relevant information on the web about a particular product.
- Travel recommendation, where an agent reads chats about travel and finds relevant information on the travel websites about a hotel or an activity.
- Facebook recommendation, where an agent reads wall postings and chats, and finds a piece of relevant information for friends on the web.

Search evaluation environment

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- For an individual query, the relevance was estimated as a percentage of correct hits among the first thirty, using the values: {correct, marginally correct, incorrect}.
- Accuracy of a single search session is calculated as the percentage of correct search results plus half of the percentage of marginally correct search results.
- Accuracy of a particular search setting (query type and search engine type) is calculated, averaging through 40 search sessions.
- For each search session, we re-ranked first 20 answers to assess relevance of PT-supported search.

Search environment

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- For our evaluation, we used customers' queries to eBay entertainment (3000 event descriptions per day are processed, PTs built and indexed). Also, eBay product-related domains is used (>200.000.000 products are searchable at one time).
- Entertainment domain of eBay (StubHub.com) is served by SOLR. The size of search index for the currently available events is 2G.
- Queries range from simple questions referring to a particular product a particular user needs, as well as a multi-sentence forum-style request to share a recommendation.
- The totality of queries was split into noun-phrase class, verb-phrase class, how-to class, and also independently split in accordance to query length (from 3 keywords to multiple sentences).

Query type	Query complexity	Relevance of baseline Bing search, %, averaging over 100 searches	Relevance of single-sentence phrase-based generalization search, %, averaging over 100 searches	Relevance of thickset-based phrase generalization search, %, averaging over 100 searches	Relevance of parse thickset-based graph generalization search, %, averaging over 100 searches
Product recommendation search	1compound sentence	62.3	69.1	72.4	73.3
	2 sent	61.5	70.5	71.9	71.6
	3 sent	59.9	66.2	72	71.4
	4 sent	60.4	66	68.5	66.7
Travel recommendation search	1compound sent	64.8	68	72.6	74.2
	2 sent	60.6	65.8	73.1	73.5
	3 sent	62.3	66.1	70.9	72.9
	4 sent	58.7	65.9	72.5	71.7
Facebook friend agent support search	1compound sent	54.5	63.2	65.3	67.2
	2 sent	52.3	60.9	62.1	63.9
	3 sent	49.7	57	61.7	61.9
	4 sent	50.9	58.3	62	62.7
Average		58.15	64.75	68.75	69.25

Search relevance was compared to Bing/Yahoo APIs as a baseline, and pair-wise sentence-sentence similarity

Data & Performance

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Mostly used entertainment products and entertainment content (data on performers, social data, search queries). Covered by distributed SOLR.
- In an industrial search application where phrases are stored in an inverse index, the generalization operation can be completed in constant time, irrespectively of the size of index. A special search request handler implementing PT matching is provided for SOLR.
- Although indexing for PTs versus regular text is much slower, it is done using distributed framework.

Search relevance improvement

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Unfiltered precision is 58.2%,
- Improvement by pair-wise sentence generalization is 11%,
- Thicket phrases – additional 6%,
- Taking projections to set of phrases reduces precision by 0.5%,
- The higher the complexity of sentence, the higher the contribution of generalization technology, from sentence level to thicket phrases to graphs.

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Узорные структуры для представления неточности

Similarity Operation for Intervals

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Similarity operation \sqcap as an interval algebra

Given $a_1, b_1, a_2, b_2 \in \mathbb{R}$,

$$[a_1, b_1] \sqcap [a_2, b_2] = [\min(a_1, a_2), \max(b_1, b_2)]$$

Intuitively, the similarity \sqcap of two intervals is the smallest interval “containing” them.

Example

$$[4, 5] \sqcap [5, 8] = [4, 8]$$

$$[3, 4] \sqcap [1, 2] = [1, 4]$$

\sqcap is idempotent, commutative and associative.

Interval Ordering

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Partial order \sqsubseteq on intervals

Given two intervals $i_1 = [a_1, b_1]$ and $i_2 = [a_2, b_2]$, the order on them is given by: $i_1 \sqsubseteq i_2$

$$\Leftrightarrow i_1 \cap i_2 = i_1$$

$$\Leftrightarrow [a_1, b_1] \cap [a_2, b_2] = [a_1, b_1]$$

$$\Leftrightarrow a_1 \leq a_2 \text{ and } b_1 \geq b_2$$

Intuitively, smaller intervals subsume larger intervals
“containing” them.

Example

$$[4, 8] \sqsubseteq [6, 8] \text{ as } 4 \leq 6 \text{ and } 8 \geq 8$$

$$[2, 5] \not\sqsubseteq [1, 8] \text{ as } 2 \not\leq 1 \text{ or } 5 \not\geq 8$$

Interval Patterns

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Interval Patterns

An *interval pattern* e is a p -dimensional vector of intervals.

$$e = \langle [a_i, b_i] \rangle_{i \in [1, p]}$$

Example

For gene expression data (GED) each gene expression profile is an interval pattern [M.Kaytoue et al., 2009-2013], e.g.

$$\langle [5, 5], [7, 7], [6, 6] \rangle, \text{ with } p = 3,$$

where $[a, a]$ stays for any number a .

Interval pattern structure $(G, (D, \delta))$. An example

Example

	s_1	s_2	s_3
g_1	5	7	6
g_2	6	8	4
g_3	4	8	5
g_4	4	9	8
g_5	5	8	5

$$G = \{g_1, \dots, g_5\}$$

$$\delta(g_1) = \langle [5, 5], [7, 7], [6, 6] \rangle$$

$$D = \{\delta(g_1), \dots, \delta(g_5)\}$$

(D, \sqcap) is a meet-semi lattice of interval patterns.

Pattern Concept. An example

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

	s_1	s_2	s_3
g_1	5	7	6
g_2	6	8	4
g_3	4	8	5
g_4	4	9	8
g_5	5	8	5

$$\{g_4, g_5\}^\square = \bigcap_{g \in \{g_4, g_5\}} \delta(g)$$

$$\{g_4, g_5\}^\square = \delta(g_4) \cap \delta(g_5)$$

$$\{g_4, g_5\}^\square = \langle [4, 4], [9, 9], [8, 8] \rangle \cap \langle [5, 5], [8, 8], [5, 5] \rangle$$

$$\{g_4, g_5\}^\square = \langle [4, 4] \cap [5, 5], [9, 9] \cap [8, 8], [8, 8] \cap [5, 5] \rangle$$

$$\{g_4, g_5\}^\square = \langle [4, 5], [8, 9], [5, 8] \rangle$$

$$\{[4, 5], [8, 9], [5, 8]\}^\square = \{g \in G \mid \langle [4, 5], [8, 9], [5, 8] \rangle \subseteq \delta(g)\}$$

$$\{[4, 5], [8, 9], [5, 8]\}^\square = \{g_3, g_4, g_5\}$$

The pair $(\{g_3, g_4, g_5\}, \langle [4, 5], [8, 9], [5, 8] \rangle)$ is a pattern concept.

Pattern concept lattice. An example

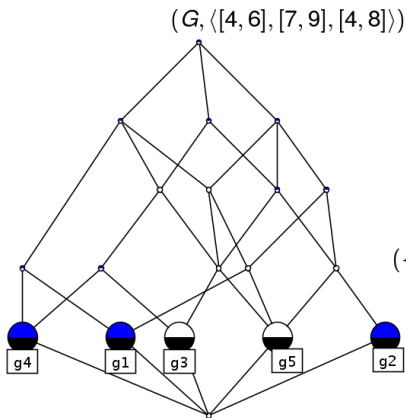
Решетки
формаль-
ных поня-
тий в со-
времен-
ных мето-
дах ана-
лиза
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их приложения

Узорные
структуры и
вызов
больших
данных



	s_1	s_2	s_3
g_1	5	7	6
g_2	6	8	4
g_3	4	8	5
g_4	4	9	8
g_5	5	8	5

$(\{g_2, g_5\}, \langle [5, 6], [8, 8], [4, 5] \rangle)$

\cup

\cap

$(\{g_2\}, \langle [6, 6], [8, 8], [4, 4] \rangle)$

Interestingness of an interval pattern

Not all pattern concepts are interesting

- Biologists look for homogeneous groups of genes: concepts having an interval pattern with “small” intervals.
- E.g., Top concept is composed of largest intervals.

A solution: introduce a *max_size* parameter

Given $d = \langle [a_i, b_i] \rangle_{i \in [1, p]}$, two constraints are defined:

- $\exists i \in [1, p] \ (b_i - a_i) \leq \text{max_size}$
- $\forall i \in [1, p] \ (b_i - a_i) \leq \text{max_size}.$

Another solution

When computing \sqcap , replace any $[a, b]$ with $b - a > \text{max_size}$ by a $*$ -value. Then, $* \sqsubseteq [a, b] \Leftrightarrow * \sqcap [a, b] = *$ for any $[a, b]$.

Knowledge Discovery with interval pattern structures

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

As for pattern structures in general, for interval pattern structures, representing imprecise data, one can define

- concepts as clusters and concept lattices as taxonomies of the field
- implication and hypotheses as strict dependencies in data
- association rules as probabilistic dependencies in data

Узорные структуры и вызов больших данных

Complexity of Computing Minimal Implication Bases

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Recall that pseudo-intents are premises of the cardinality-minimal implication base.

- The number of pseudo-intents can be exponential, the problem of counting pseudo-intents is $\#P$ -hard. [Kuznetsov 2004]
- Pseudo-intents cannot be generated in lexicographic order with polynomial delay unless $P=NP$, generating pseudo-intents is TRANSENUM-hard [F.Distel, B.Sertkaya 2010]
- Recognizing whether a subset of attributes is a pseudo-intent is coNP-complete [Kuznetsov, Obiedkov 2006], [Babin, Kuznetsov 2010]
- Efficient parallelization approaches are not known

"Knowledge" is hard to compute and can be much larger than data!
For some tasks data may be a better "basis"

Lazy evaluation vs. computing bases

If you need implications for classification, make classification directly from data

If you need "understanding" data, compute a subbase: small-premise implications or those implicitly used for classification (anecdotal base)

Short-Premise Base

$a \rightarrow \dots$
 $ab \rightarrow \dots$
 $abc \rightarrow \dots$
 $abcd \rightarrow \dots$

g_1
g_2
g_k

→ Implication Base

Lazy Classification

Classification

g_{new}	?
------------------	---

Good News: Classification with $(\cdot)''$ -closure

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

For an arbitrary subset of attributes $A \subseteq M$ the maximal set which can be deduced with implications of the context is A''

What is A'' ? Take all objects that contain A and intersect them.

This takes $O(|G| \cdot |M|)$ time for binary data tables and $O(|G| \cdot p(\sqsubseteq) + |G| \cdot p(\sqcap)) = O(|G| \cdot p(\sqcap))$ time for pattern structures, where $p(\sqsubseteq)$ and $p(\sqcap)$ are times for computing \sqsubseteq and \sqcap , respectively.

Lazy classification: An example

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

$G \setminus M$	target	m_1	m_2	m_3	m_4	m_5	m_6
g_1			×	×		×	×
g_2		×		×	×		×
g_3		×	×		×	×	
g_4	×		×	×	×	×	×
g_5	×	×		×	×	×	×
g_6	×	×	×		×	×	×
g_7	×	×	×	×		×	×
g_8	×	×	×	×	×		×
g_9	×	×	×	×	×	×	

$2^{|M|/2} = 8$ implications in the minimal base, with the premises

$\{m_1, m_2, m_3\}, \{m_1, m_2, m_6\}, \{m_1, m_5, m_3\}, \{m_1, m_5, m_6\},$

$\{m_4, m_2, m_3\}, \{m_4, m_2, m_6\}, \{m_4, m_5, m_3\}, \{m_4, m_5, m_6\}.$

To classify a new object g w.r.t. target t , compute $(g' \cap g_i')''$, which takes $O(|G|^2 \cdot |M|)$ time.

If $g' = \{m_1, m_2, m_5\}$, then for all g_i one has
 $t \not\subseteq (g'_i \cap \{m_1, m_2, m_5\})''$, hence g is classified negatively.

Lazy classification: An example

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

$G \setminus M$	target	m_1	m_2	m_3	m_4	m_5	m_6
g_1			×	×		×	×
g_2		×		×	×		×
g_3		×	×		×	×	
g_4	×		×	×	×	×	×
g_5	×	×		×	×	×	×
g_6	×	×	×		×	×	×
g_7	×	×	×	×		×	×
g_8	×	×	×	×	×		×
g_9	×	×	×	×	×	×	

$2^{|M|/2} = 8$ implications in the minimal base, with the premises

$\{m_1, m_2, m_3\}, \{m_1, m_2, m_6\}, \{m_1, m_5, m_3\}, \{m_1, m_5, m_6\},$

$\{m_4, m_2, m_3\}, \{m_4, m_2, m_6\}, \{m_4, m_5, m_3\}, \{m_4, m_5, m_6\}.$

To classify a new object h w.r.t. target t , compute $(h' \cap g_i'')''$, which takes $O(|G|^2 \cdot |M|)$ time. If $h' = \{m_1, m_2, m_3, \}$, then

for g_7 one has $(g_7 \cap \{m_1, m_2, m_3\})'' = \{m_1, m_2, m_3, t\}$, hence h is classified positively.

Classification with Hypotheses in $O(|G|^2 \cdot |M|)$ time

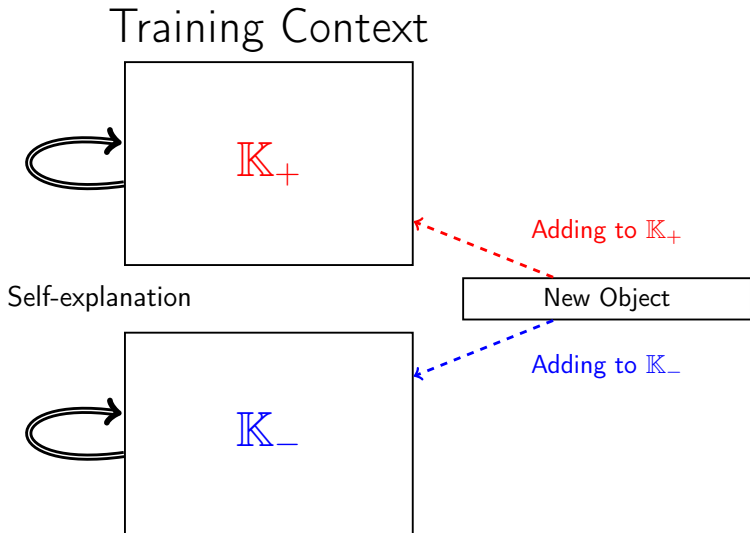
Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

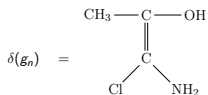
АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

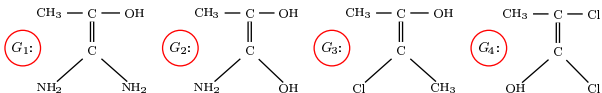
Узорные
структуры и
вызов
больших
данных

Ленивая классификация с использованием импликаций

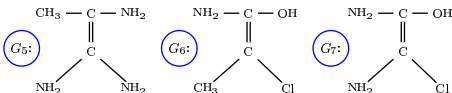
How a new example g_n is classified?



Positive examples:



Negative examples:



Positive lattice

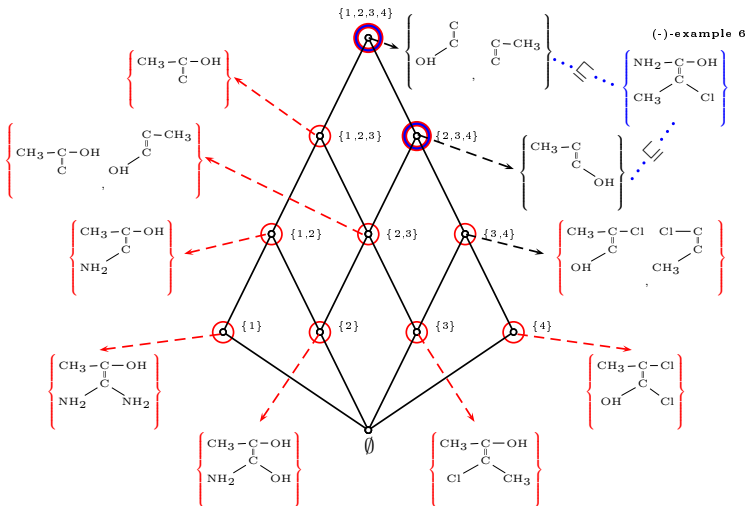
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Using $(\cdot)^\diamond$ for lazy evaluation: positive classification

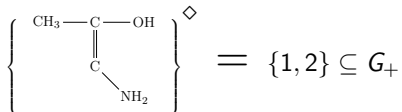
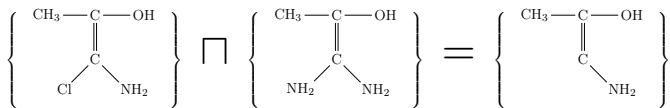
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
большой
данных



This takes $O(|G| \cdot (p(\sqcap) + |G| \cdot p(\sqsubseteq)))$ time, where $p(\sqsubseteq)$ and $p(\sqcap)$ are times for computing \sqsubseteq and \sqcap , respectively.

For projections of fixed size where $p(\sqcap) = O(1)$ classification of a new object takes $O(|G|^2)$ time.

Using $(\cdot)^\diamond$ for lazy evaluation: negative classification of g_m

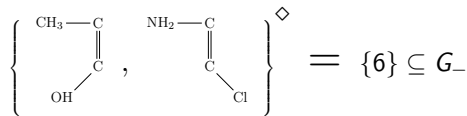
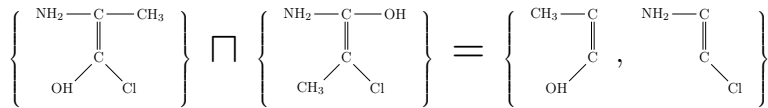
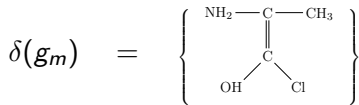
Решетки
формальных
понятий
в современных
методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных



Parallelization and Randomization of Lazy Classification with Implications

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Computing closure is trivially parallelizable: one partitions the set of objects, computes closures in each part, and intersects all the partial closures. So, classifying an object with implications using k processors takes $O(|G| \cdot (|G| \cdot p(\sqsubseteq) + p(\sqcap)) \cdot 1/k)$ time;
- If one allows for certain tolerance in classification, the approach becomes easily randomizable: one can apply usual Monte Carlo techniques for estimating classification: one chooses a sample from positive and negative examples and computes closures with respect to them, thus obtaining an estimate for probabilities of classification with complete data.

For projections of fixed size where $p(\sqcap) = O(1)$ classification of a new object takes $O(|G|^2/k)$ time.

Выводы

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Анализ формальных понятия и узорные структуры могут эффективно использоваться для анализа больших сложных данных и знаний
- Сложность вычислений в узорных структурах с применением проекций, ленивой классификации, параллелизации и рандомизации может быть полиномом низкой степени, что позволяет применять эти модели для анализа больших данных

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

Спасибо! Вопросы?

References

- Kuznetsov, S.O.: Mathematical aspects of concept analysis. J.Math. Sci., vol. 80(2), 1654-1698 (1996).
- Ganter, B., Kuznetsov, S.O.: *Pattern Structures and Their Projections*. In: Proc. ICCS 2001. LNAI, vol. 2120, 129-142. (2001)
- Kuznetsov, S.O.: *Complexity of Learning in Concept Lattices from Positive and Negative Examples*. Discr. Appl. Math. vol. 142, 111-125 (2004)
- Kuznetsov, S.O., Samokhin, M.V.: *Learning Closed Sets of Labeled Graphs for Chemical Applications*. In: Proc. ILP 2005. LNAI, vol. 3625, 190-208. (2005)
- Kuznetsov, S.O., Obiedkov, S.A.: *Some Decision and Counting Problems of the Duquenne-Guigues Basis of Implications*. Discr. Appl. Math., vol. 156, no. 11, 1994-2003 (2008)

References

Решетки
формаль-
ных понятий
в современ-
ных методах
анализа
данных и
знаний

С.О.Кузнецов

АФП для
поиска
знаний в
данных

Узорные
структуры и
их
приложения

Узорные
структуры и
вызов
больших
данных

- Kuznetsov, S.O.: *Pattern Structures for Analyzing Complex Data*. In: Proc. RSFDGrC 2009. LNAI, vol. 5908, 33–44. (2009)
- Kuznetsov, S.O.: *Computing Graph-Based Lattices from Smallest Projections*. In: Proc. KPP 2007. LNAI, vol. 6581, 35–47. (2011)
- Kuznetsov, S.O.: *Fitting Pattern Structures to Knowledge Discovery in Big Data*. In: Proc. ICFCA 2013. LNAI vol. 7880, 254 -266. (2013)