Krajca P., Outrata J., Vychodil V.: Computing Formal Concepts by Attribute Sorting

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

:: Formal concept analysis (FCA) is a method of tabular data analysis

:: used for data mining, knowledge discovery, information retrieval, data preprocessing

:: input – object-attribute table, i.e., two dimensional table with rows representing objects, columns their attributes (features), where crosses indicate that particular object has particular attribute

	needs water	lives in water	lives on land	has chlorophyll	can move around
dog	×		×		×
dolphin	×	×			×
frog	×	×	×		×
bean	×		×	×	
daffodil	×		×	×	
waterlily	×	×		×	

Table 1.: Formal context

:: output – all maximal submatrices full of $\times$'s present in table

- these submatrices are natural concepts hidden in the data (e.g., animal, fish)
- form a hierarchy (e.g., fish $\leq$ animal)

Definitions

:: **Formal context** $\mathbb{K}$ is a triplet $\langle X, Y, I\rangle$, where X and Y are non-empty sets and $I \subseteq X \times Y$. (X ...set of objects, Y ...set of attributes, $\langle x, y\rangle \in I$ ...object x has attribute y)

:: **Concept-forming operators**: For a formal context $\mathbb{K} = \langle X, Y, I\rangle$, operators ${}^{\uparrow_\mathbb{K}} : 2^X \to 2^Y$ and ${}^{\downarrow_\mathbb{K}} : 2^Y \to 2^X$ are defined for every $A \subseteq X$ and $B \subseteq Y$ by:

$$A^{\uparrow_\mathbb{K}} = \{y \in Y \mid \text{for each } x \in A: \ \langle x, y\rangle \in I\},$$
$$B^{\downarrow_\mathbb{K}} = \{x \in X \mid \text{for each } y \in B: \ \langle x, y\rangle \in I\}.$$

$A^{\uparrow_\mathbb{K}}$ ...set of all attributes shared by all objects from A
$B^{\downarrow_\mathbb{K}}$ ...set of all objects sharing all attributes from B

:: **Formal concept** in $\mathbb{K} = \langle X, Y, I\rangle$ is a pair $\langle A, B\rangle$ of $A \subseteq X$ and $B \subseteq Y$ such that $A^{\uparrow_\mathbb{K}} = B$ and $B^{\downarrow_\mathbb{K}} = A$.

Preliminaries

R-context

:: formal context derived from $\mathbb{K} = \langle X, Y, I\rangle$

:: attributes are pairs $\langle flag, B\rangle$ where B is a subset of Y, $flag \in \mathbb{N}_0$

Definition 1. Given a formal context $\mathbb{K} = \langle X, Y, I\rangle$, a triplet $\mathbb{K}^\sharp = \langle X^\sharp, Y^\sharp, I^\sharp\rangle$ is called an *R-context* (*derived from* $\mathbb{K}$) if the following conditions are satisfied:

(1) $X^\sharp \subseteq X$;
(2) $Y^\sharp \subseteq \mathbb{N}_0 \times 2^Y$ such that for any $\langle n_1, B_1\rangle \in Y$ and $\langle n_2, B_2\rangle \in Y$ we have either that (a) $n_1 = n_2$ and $B_1 = B_2 \neq \emptyset$ or (b) $B_1 \neq \emptyset$, $B_2 \neq \emptyset$, and $B_1 \cap B_2 = \emptyset$;
(3) for any $x \in X^\sharp$ and $\langle n, B\rangle \in Y^\sharp$: $\langle x, y_1\rangle \in I$ iff $\langle x, y_2\rangle \in I$ holds true for all $y_1, y_2 \in B$;
(4) $I^\sharp = \{\langle x, \langle n, B\rangle\rangle \in X^\sharp \times Y^\sharp \mid \langle x, y\rangle \in I$ for all $y \in B\}$.

Auxiliary definitions

:: R-context derived from $\mathbb{K}$ is called *initial* if $X^\sharp = X$, $Y^\sharp = \{\langle 0, \{y\}\rangle \mid y \in Y\}$, and $I^\sharp = \{\langle x, \langle 0, \{y\}\rangle\rangle \in X^\sharp \times Y^\sharp \mid \langle x, y\rangle \in I\}$

:: $\lfloor D\rfloor = \bigcup\{B \subseteq Y \mid \langle n, B\rangle \in D\}$, for any $D \subseteq Y^\sharp$

:: $\textsc{Int}(\mathbb{K}^\sharp, Y) = Y \setminus \lfloor Y^\sharp\rfloor$, for any $\mathbb{K}^\sharp$ derived from formal context $\langle X, Y, I\rangle$

Clarification

:: formal context $\mathbb{K} = \langle X, Y, I\rangle$ is called *clarified* if for any $y_1, y_2 \in Y$ it follows that $\{y_1\}^\downarrow = \{y_2\}^\downarrow$ implies $y_1 = y_2$ and dually for any couple of objects

Definition 2. For any R-context $\mathbb{K}^\sharp = \langle X^\sharp, Y^\sharp, I^\sharp\rangle$, we define clarified context $\mathbb{K}^{\mathsf{C}}$ as a triplet $\langle X^{\mathsf{C}}, Y^{\mathsf{C}}, I^{\mathsf{C}}\rangle$ where

(1) $X^{\mathsf{C}} = X^\sharp$;
(2) $Y^{\mathsf{C}} = \left\{\left\langle \sum\{n \in \mathbb{N}_0 \mid \langle n, B\rangle \in [y]_{\equiv_{\mathbb{K}^\sharp}}\}, \left\lfloor [y]_{\equiv_{\mathbb{K}^\sharp}}\right\rfloor \right\rangle \mid y \in Y^\sharp\right\}$,
(3) $I^{\mathsf{C}} = \left\{\langle x, \langle n, B\rangle\rangle \in X^{\mathsf{C}} \times Y^{\mathsf{C}} \mid \right.$ there is $n' \leq n$ and $B' \subseteq B$ such that $\left.\langle x, \langle n', B'\rangle\rangle \in I^\sharp\right\}$.

$\mathbb{K}^{\mathsf{C}}$	$\langle 0, \{1, 4\}\rangle$	$\langle 1, \{2, 7\}\rangle$	$\langle 0, \{3\}\rangle$	$\langle 0, \{6\}\rangle$
b		×	×	
d	×	×		×
e	×			
f		×	×	

Table 2.: Example of clarified R-context

Algorithm

:: algorithms (Ganter, CbO, FCbO) compute some concepts multiple times (significant overhead)

:: observation: reordering of attributes according to their support reduces number of multiple times computed concepts

:: we consider a bijective map $f: Y^\sharp \to \{0, \ldots, |Y^\sharp| - 1\}$ such that, for any $y_1, y_2 \in Y^\sharp$,

$$\text{if } f(y_1) \leq f(y_2) \text{ then } |\{y_1\}^{\downarrow_{\mathbb{K}^\sharp}}| \leq |\{y_2\}^{\downarrow_{\mathbb{K}^\sharp}}|.$$

:: f represents a position in an ordered list of attributes which are sorted according to their support

:: map f along with $flag$ plays the role of the canonicity test ensuring that each concept is returned only once

Sketch of the Algorithm

:: algorithm (procedure $\textsc{Compute}$) starts with a clarified initial R-context

:: computes closure for each attribute having zero flag

:: if closure does not contain any attribute with nonzero flag, it is used to derive new R-context

:: new context is clarified and flags are updated

:: algorithm recursively applies procedure $\textsc{Compute}$

Procedure $\textsc{Compute}(\mathbb{K}^\sharp)$
store $\langle X^\sharp, \textsc{Int}(\mathbb{K}^\sharp, Y)\rangle$
for $\langle n, B\rangle \in Y^\sharp$ **do**
  if $n = 0$ **then**
    set $\langle C, D\rangle$ **to** $\langle\langle n, B\rangle^{\downarrow_{\mathbb{K}^\sharp}}, \langle n, B\rangle^{\downarrow\uparrow_{\mathbb{K}^\sharp}}\rangle$
    if $\sum\{n \in \mathbb{N}_0 \mid \langle n, B\rangle \in D\} = 0$ **then**
      $\textsc{Compute}(\textsc{Reduce}(\mathbb{K}^\sharp, C, D))$

:: $\textsc{Reduce}(\mathbb{K}^\sharp, C, D)$ – returns clarified R-context such that

(1) $X^{\mathfrak{R}} = C$;
(2) $Y^{\mathfrak{R}} = \{\mathrm{Attr}(y) \mid y \in Y^\sharp$ and $y \notin D\}$, where $\mathrm{Attr}(y) \in \mathbb{N}_0 \times 2^Y$ is defined by

$$\mathrm{Attr}(\langle n, B\rangle) = \begin{cases} \langle |B|, B\rangle, & \text{if } n = 0 \text{ and } f(\langle n, B\rangle) < f(\min(D)), \\ \langle n, B\rangle, & \text{otherwise.} \end{cases}$$

(3) $I^{\mathfrak{R}} = \left\{\langle x, \langle n, B\rangle\rangle \in X^{\mathfrak{R}} \times Y^{\mathfrak{R}} \mid \right.$ there is $n' \leq n$ such that $\left.\langle x, \langle n', B\rangle\rangle \in I^\sharp\right\}$.

:: attribute $\langle 0, B\rangle \in Y^\sharp$ will be given a nonzero flag in $Y^{\mathfrak{R}}$ if it is not in D and if it stays before $\min(D)$ in terms of the order of attributes

Complexity and Efficiency

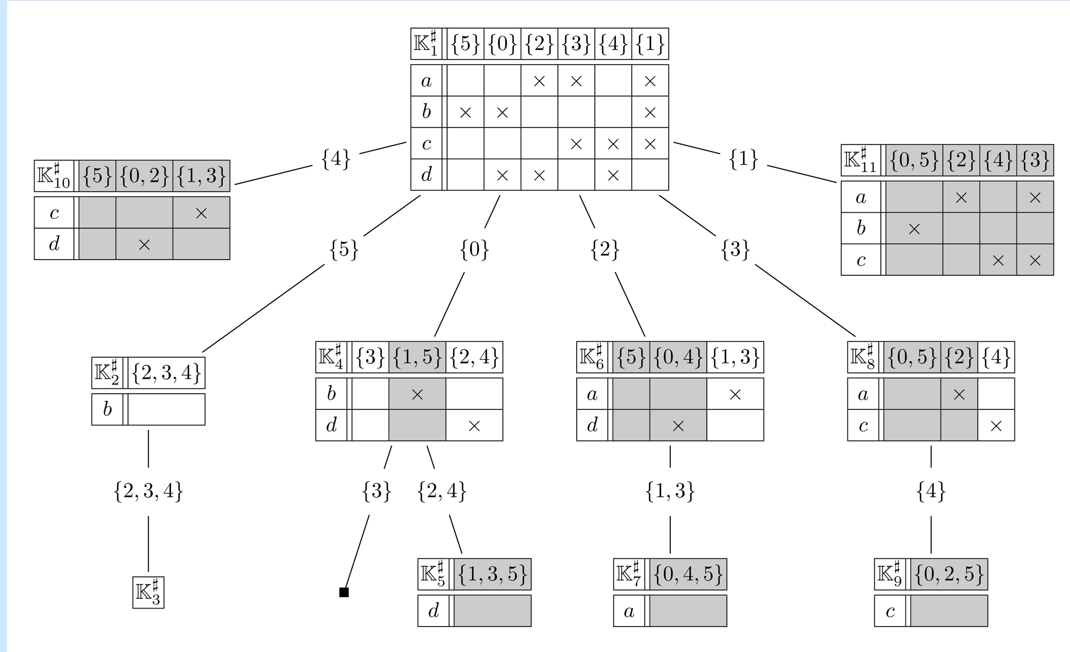

Figure 1.: Invocations of the Compute procedure

Complexity

:: complexity: $O(|\mathcal{B}(X, Y, I)| \cdot |X| \cdot |Y|^2)$, where $\mathcal{B}(X, Y, I)$ is a set of all formal concepts in a formal context $\langle X, Y, I\rangle$

:: polynomial time delay: $O(|Y|^3 \cdot |X|)$

Comparisons

	debian tags	anon. web.	mushroom
size	$14,315 \times 475$	$32,710 \times 295$	$8,124 \times 119$
density	$< 1\,\%$	$1\,\%$	$19\,\%$
# concepts	38,977	129,009	238,710
Attr. sort.	44,221	135,925	246,181
FCbO (ord.)	298,641	398,147	299,201
FCbO	679,911	1,475,341	426,563
CbO (ord.)	960,106	785,394	1,321,524
CbO	12,045,680	27,949,552	4,006,498

Table 3.: Number of closures computed by selected algorithms from CbO family

	mean value	std. dev.	median value
CbO	3,359.88	505.51	3294
CbO (ord.)	1,394.08	78.19	1,395
FCbO	860.41	49.17	860
FCbO (ord.)	853.87	47.80	852
Attr. sort.	240.83	8.34	241
# concepts	227.58	6.79	228

Table 4.: Computed closures in datasets 50×50 with 10 % density of 1's

International Center for Information and Uncertainty

SSIU 2012: International Summer School "Information and Uncertainty"

http://mcin.upol.cz/SSWIU-2012/

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness