Rank-aware Clustering of Relational Data: Organizing Search Results

Petr Krajča

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY, OLOMOUC

Motivation

- applications of similarity-based databases
- improving user experience

#	*Score*	*City*	*Price*	*Bdrms*	*SqFeet*	*Porch*
1	1.000	Roseville	327,000	5	3,856	Y
2	0.850	Roseville	321,900	5	4,460	Y
3	0.560	Elmwood	290,000	5	2,933	N
4	0.560	West End	292,000	3	2,945	Y
5	0.560	Roseville	295,900	5	3,820	Y
6	0.325	West End	299,900	3	2,810	N
7	0.275	Roseville	181,500	4	2,562	Y

Issues

- users overwhelmed with similar items
- items with similar relevance (score) mixed up (unintuitive order)
- lack of insight into result ordering

Proposed Solution

- based on formal concepts analysis (FCA)

Outline of the Algorithm

1. convert input data into a form suitable for FCA
2. identify formal concepts (clusters)
3. from these concepts pick the most interesting ones from the user's viewpoint

Remarks

- FCA: well-established framework (theory, algorithms, applications)
- connection to psychology of concepts
- need to preserve order given by the *scoring* function

Formal Concept Analysis (1 of 3)

- method of tabular data analysis (R. Wille, TU Darmstadt)
- used for data mining, knowledge discovery, data preprocessing

Input

- table—rows = objects, columns = attributes (features), $\times$ indicates that particular object has particular attribute

	a_1	a_2	a_3	a_4
o_1	$\times$	$\times$		$\times$
o_2	$\times$		$\times$	
o_3		$\times$	$\times$	$\times$
o_4	$\times$	$\times$	$\times$	$\times$

Output

- all maximal submatrices full of $\times$'s present in table
- these submatrices are natural concepts hidden in the data
- form a hierarchy

Formal Concept Analysis (2 of 3)

A **formal context** 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 $\langle X, Y, I \rangle$, operators ${}^{\uparrow} : 2^X \to 2^Y$ and ${}^{\downarrow} : 2^Y \to 2^X$ are defined for every $A \subseteq X$ and $B \subseteq Y$ by:

$$A^{\uparrow} = \{y \in Y \mid \text{for each } x \in A : \ \langle x, y \rangle \in I\},$$
$$B^{\downarrow} = \{x \in X \mid \text{for each } y \in B : \ \langle x, y \rangle \in I\}.$$

- $A^{\uparrow}$ ... set of all attributes shared by all objects from A
- $B^{\downarrow}$ ... set of all objects sharing all attributes from B

Formal Concept Analysis (3 of 3)

A **formal concept** in $\langle X, Y, I\rangle$ is a pair $\langle A, B\rangle$ of $A \subseteq X$ and $B \subseteq Y$ such that

$$A^{\uparrow} = B \text{ and } B^{\downarrow} = A.$$

- A . . . extent of $\langle A, B\rangle$
- B . . . intent of $\langle A, B\rangle$
- $\langle A, B\rangle$ is a formal concept iff A contains just objects sharing all attributes from B and B contains just attributes shared by all objects from A.

Formal Concept Analysis (Example)

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

$$\{dog, cod, frog\}^{\uparrow} = \{needs\ water, can\ move\ around\}$$

$$\{needs\ water, can\ move\ around\}^{\downarrow} = \{dog, cod, frog\}$$

$\langle\{dog, cod, frog\}, \{needs\ water, can\ move\ around\}\rangle \Longrightarrow$ animal

Subconcept-superconcept Hierarchy

- partial order $\leq$

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \text{ iff } A_1 \subseteq A_2 \text{ (or, equivalently, iff } B_2 \subseteq B_1).$$

- set of formal concepts $\mathcal{B}(X, Y, I)$ together with $\leq$ form a complete lattice (concept lattice).

Natural interpretation

- *animal*: $\langle\{dog, cod, frog\}, \{needs\ water, can\ move\ around\}\rangle$
- *dog*: $\langle\{dog\}, \{needs\ water, lives\ on\ land, can\ move\ around\}\rangle$
- $dog \leq animal$, this means:
  - *dog* – more specific concept
  - *animal* – more general concept

Input Data

- ranked data table
- $\mathbb{Y} = \{y_1, \dots, y_n\}$ finite number of columns (attributes)
- each attribute has its domain D_y (set of permitted values)
- **Cartesian product of domains**, denoted by $\prod_{y \in \mathbb{Y}} D_y$, is a set of all maps

$$t \colon \mathbb{Y} \to \bigcup_{y \in \mathbb{Y}} D_y$$

  such that $t(y) \in D_y$ for all $y \in \mathbb{Y}$.
- **data table** is any finite subset $\mathcal{D} \subseteq \prod_{y \in \mathbb{Y}} D_y$.
- $\mathcal{D}$ is a set of tuples (no inherent order of tuples)
- let $\langle \mathbb{S}, \leq \rangle$ be a poset, map $s_{\mathcal{D}}$

$$s_{\mathcal{D}} : \mathcal{D} \to \mathbb{S}$$

  describes relevance of tuples in the data table (scoring function)

Data Preparation (1 of 2)

- **conceptual scaling** is a process transforming general data table $\mathcal{D}$ into a formal context $\langle X, Y, I\rangle$
- replacing ordinal attributes with nominal ones (e.g., with equidistant intervals)
- e.g.: D_{price} may be replaced with intervals $\{\ldots, [280,000; 290,000), [290,000; 300,000), \ldots\}$
- $X = \{1, \ldots, n\}$ where each $x \in X$ corresponds to one row t in the data table and numbers are assigned to rows in the descending order w.r.t. $s_{\mathcal{D}}$
- $Y = \left\{\langle y, v\rangle \mid \langle y, v\rangle \in \bigcup_{t_i \in \mathcal{D}} t_i\right\}$, i.e., all attribute value pairs in the data table $\mathcal{D}$
- $I = \{\langle i, \langle y, v\rangle\rangle \mid \text{for every } t_i \in \mathcal{D} \text{ and } y \in Y \text{ iff } t_i(y) = v\}$ (object i has an attribute $\langle y, v\rangle$, iff the value of the attribute y of row t_i is equal to v)

Data Preparation (2 of 2)

- map $r : X \to \mathbb{N}$ assigns to each tuple numerical rank such that for every two tuples $t_i, t_j \in \mathcal{D}$ and corresponding objects $x_i, x_j \in X$,

$$s_{\mathcal{D}}(x_i) \leq s_{\mathcal{D}}(x_j) \text{ implies } r(x_j) \leq r(x_i).$$

- r and $\leq$ provides comparative meaning
- $r(x_i) \leq r(x_j)$ means object x_i is more or equally relevant than x_j

Formal Context for Our Running Example

x	$r(x)$	$\langle City, Roseville \rangle$	$\langle City, Elmwood \rangle$	$\langle City, WestEnd \rangle$	$\langle Price, 320k \rangle$	$\langle Price, 290k \rangle$	$\langle Price, 180k \rangle$	$\langle Bdrms, 3 \rangle$	$\langle Bdrms, 4 \rangle$	$\langle Bdrms, 5 \rangle$	$\langle SqFeet, 2.4k \rangle$	$\langle SqFeet, 2.8k \rangle$	$\langle SqFeet, 3.8k \rangle$	$\langle SqFeet, 4.4k \rangle$	$\langle Porch, Y \rangle$	$\langle Porch, N \rangle$
1	1	×			×					×			×		×	
2	2	×			×					×				×	×	
3	5		×			×				×		×				×
4	5			×		×		×				×			×	
5	5	×				×				×			×		×	
6	6			×		×		×				×				×
7	7	×					×		×		×				×	

Algorithm: Idea (1 of 2)

- each formal concept $\langle A, B\rangle$ identifies set of objects A having common attributes B
- set of attributes B unambiguously describes set of objects A
- attributes from B can serve as description (labels) for objects from A
- interested in formal concepts creating continuous sequence w.r.t. a ranking function

Formal concept $\langle A, B\rangle$ shall be called **continuous formal concept** w.r.t. a ranking function r iff there is no object $x \notin A$ such that

$$\min_{i\in A}(r(i)) < r(x) < \max_{i\in A}(r(i)).$$

Algorithm: Idea (2 of 2)

- enumerating continuous formal concepts recursively in lexicographical order $\prec$
- e.g., $\langle\{1,2\},\ldots\rangle \prec \langle\{1,2,3\},\ldots\rangle \prec \langle\{1,3\},\ldots\rangle$
- recursive algorithm
  - each invocation extends input formal concept with one object
  - whenever is the new concept lexigraphically smaller, the given branch of computation can be abandoned
- variant of the Kuznetsov's Close-by-One (CbO) algorithm
- extension enumerating only continuous formal concepts (pruning)

Algorithm: Pseudocode

```
Procedure Generate(⟨A, B⟩, x)
    result = {⟨A, B⟩};
    foreach i in X − A do
        if (i ≥ x) and ((A ≠ ∅) and (r(i) ≤ MaxRank(A, x))) then
            D = B ∩ {i}↑;
            C = D↓;
            skip = false;
            foreach j in C do
                if ((j < i) ∧ (j ∉ A)) then
                    skip = true;
                    break loop;
            if not skip and IsContinuous(C, i) then
                result = result ∪ Generate(⟨C, D⟩, i);
    return result;
```

Algorithm 1: Procedure computing all continuous formal concepts

Are All Concepts Equal?

- large number of formal concepts hidden in the data
- not all continuous
- still large number (18 in our examples)
- some of low importance, e.g.:
  - covering single object
  - covering single attribute $\langle\{1, 2, 3, 5\}, \{\langle Bdrms, 5\rangle\}\rangle$

What Is It?

(a) a transport vehicle

(b) a car

(c) a 2011 Ford Mondeo LX Hatchback

In sentence

(a) I always go to work by transport vehicle.
(b) I always go to work by car.
(c) I always go to work by 2011 Ford Mondeo LX Hatchback.

What Is It?

(a) a transport vehicle

(b) a car

(c) a 2011 Ford Mondeo LX Hatchback

In sentence

(a) I always go to work by transport vehicle.

(b) I always go to work by car.

(c) I always go to work by 2011 Ford Mondeo LX Hatchback.

What Is It?

(a) a transport vehicle

(b) a car

(c) a 2011 Ford Mondeo LX Hatchback

In sentence

(a) I always go to work by transport vehicle.

(b) I always go to work by car.

(c) I always go to work by 2011 Ford Mondeo LX Hatchback.

Basic Level Concepts: Intuition

- multiple approaches
- we adhere to definition by E. Rosch
- notion of *cohesion* which is a measure describing similarity among objects in a given formal concept

Basic Level Concept

(a) $\langle A, B\rangle$ has a high cohesion,
(b) $\langle A, B\rangle$ has a significantly larger cohesion than its upper neighbors,
(c) $\langle A, B\rangle$ has only a slightly smaller cohesion than its lower neighbors.

- not a yes/no property

Basic Level Concepts: Formalization (1 of 2)

- approach based on fuzzy logic in the narrow sense (proposed by Belohlavek and Trnecka)

Cohesion

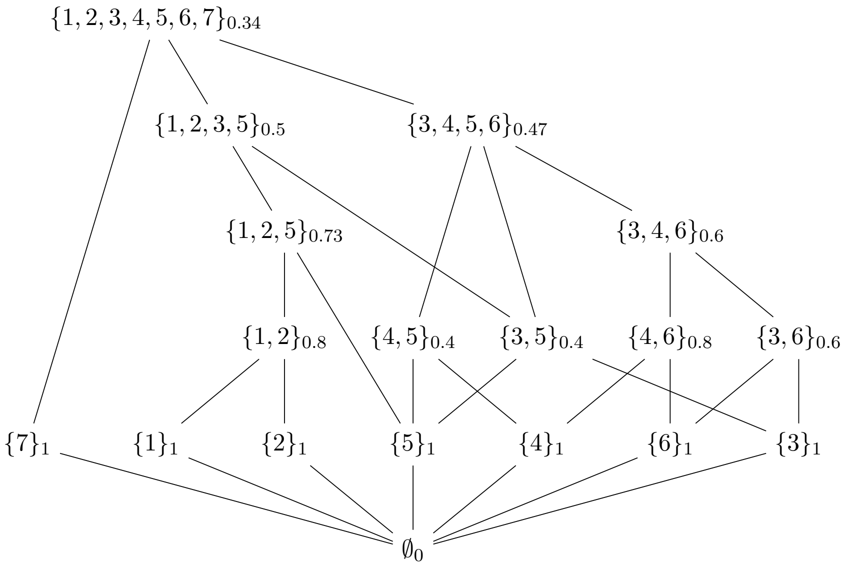
- an average bitwise similarity of all objects of a formal concept

$$coh(\langle A, B\rangle) = \frac{\sum_{x_1,x_2\in A, x_1>x_2} sim(x_1,x_2)}{|A|\cdot(|A|-1)/2}$$

- where $sim(x_1,x_2)$ is similarity of two objects, i.e., a ratio of attributes both concepts have in common to the total number of attributes

$$sim(x_1,x_2) = \frac{|\{x_1\}^{\uparrow}\cap\{x_2\}^{\uparrow}|}{|\mathbb{Y}|]}$$

Example: Continuous Formal Concepts and Cohesion



Basic Level Concepts: Formalization (2 of 2)

- formalization of properties proposed by Rosch

$$BL(c) = BL_a(c) \cdot BL_b(c) \cdot BL_c(c)$$

- real interval $[0, 1]$ as a scale of truth degrees
- multiplication corresponds to a product t-norm (Goguen)

(a) has a high cohesion $\ldots coh(c)$

(b) has a significantly larger cohesion than its UN's $\ldots 1 - \frac{coh(c_u)}{coh(c)}$ where c_u is an UN

(c) has only a slightly smaller cohesion than its LN's $\ldots \frac{coh(c)}{coh(c_l)}$ where c_l is a LN

$$BL_a(c) = coh(c)$$

$$BL_b(c) = \frac{1}{|\mathcal{UN}^*(\mathcal{B}, c)|} \cdot \sum_{c_u \in \mathcal{UN}^*(\mathcal{B}, c)} 1 - \frac{coh(c_u)}{coh(c)}$$

$$BL_c(c) = \frac{1}{|\mathcal{LN}^*(\mathcal{B}, c)|} \cdot \sum_{c_l \in \mathcal{LN}^*(\mathcal{B}, c)} \frac{coh(c)}{coh(c_l)}$$

Results: Numerical Point of View

objects	BL_a	BL_b	BL_c	BL
$\{\}$	0	1	0	0
$\{1\}$	1	0.2	0	0
$\{1,2\}$	0.8	0.08	0.8	0.05
$\{1,2,3,5\}$	0.5	0.31	0.68	0.11
$\{1,2,3,4,5,6,7\}$	0.34	0	0.59	0
$\{\mathbf{1},\mathbf{2},\mathbf{5}\}$	**0.73**	**0.32**	**0.83**	**0.19**
$\{2\}$	1	0.2	0	0
$\{\mathbf{3}\}$	**1**	**0.5**	**0**	**0**
$\{3,4,6\}$	0.6	0.22	0.88	0.12
$\{3,4,5,6\}$	0.47	0.27	0.78	0.1
$\{3,5\}$	0.4	0	0.4	0
$\{3,6\}$	0.6	0	0.6	0
$\{4\}$	1	0.4	0	0
$\{4,5\}$	0.4	0	0.4	0
$\{\mathbf{4},\mathbf{6}\}$	**0.8**	**0.25**	**0.8**	**0.16**
$\{5\}$	1	0.49	0	0
$\{6\}$	1	0.3	0	0
$\{\mathbf{7}\}$	**1**	**0.66**	**0**	**0**

Results: User-friendly Point of View

City	*Price*	*Bdrms*	*SqFeet*	*Porch*
Roseville; 5 bedrooms; Porch				
Roseville	327,000	5	3,856	Y
Roseville	321,900	5	4,460	Y
Roseville	295,900	5	3,820	Y
Elmwood	290,000	5	2,933	N
West End; $290,000; 2,800 sq. feet				
West End	292,000	3	2,945	Y
West End	299,900	3	2,810	N
Roseville	181,500	4	2,562	Y

Conclusions and Future Research

- novel efficient algorithm for organizing search engine results
- real-world issue
- takes into account psychology of concepts
- suitable for other applications
  - ordinary database query processing
  - document search engines
- large scale evaluation (incl. A/B testing)