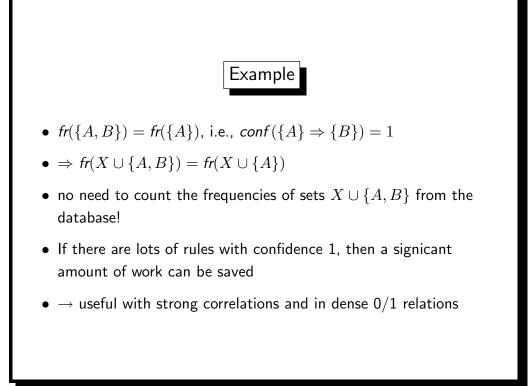
Closed sets, generators, condensed representations

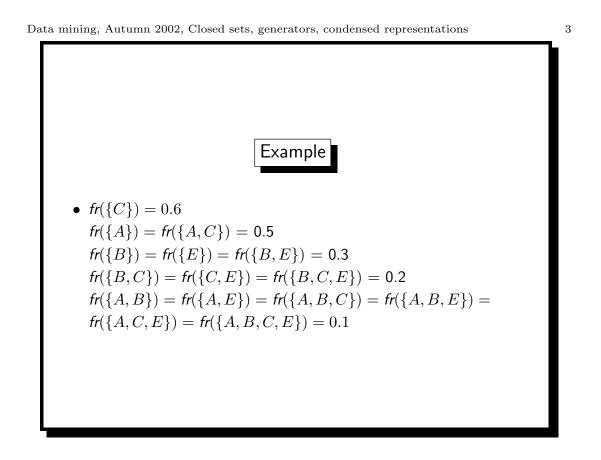
Data mining, Autumn 2002, Closed sets, generators, condensed representations

Closed sets, generators, condensed representations

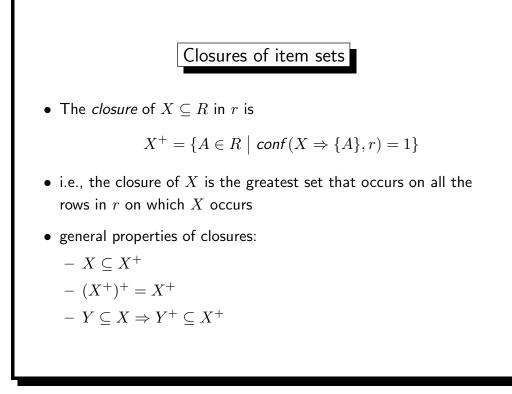
1

- Closure, closed set, generator
- Algorithms
- Condensed representations
- Experimental results
- Literature for this part





 $\mathbf{5}$



Data mining, Autumn 2002, Closed sets, generators, condensed representations

Closed sets

- item set X is *closed* iff $X^+ = X$
- the collection of all closed sets:

$$\mathcal{C}\ell = \{X^+ \mid X \subseteq R\}$$

- closed sets and their frequencies alone are a sufficient representation for the frequencies of all sets:
- either X is itself closed or some of its supersets is in any case X^+ is closed and so its frequency is known
- but which of the closed supersets of X is the closure X⁺? the one with the greatest frequency (why?)

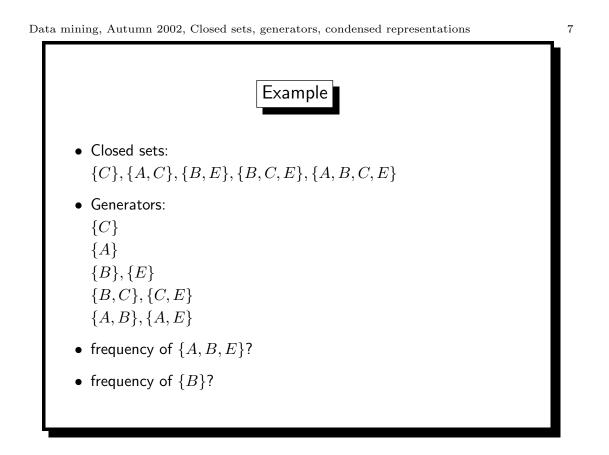
• thus:
$$fr(X) = \max\{fr(Y) \mid Y \in \mathcal{C}\ell \text{ and } X \subseteq Y\}$$

Generators

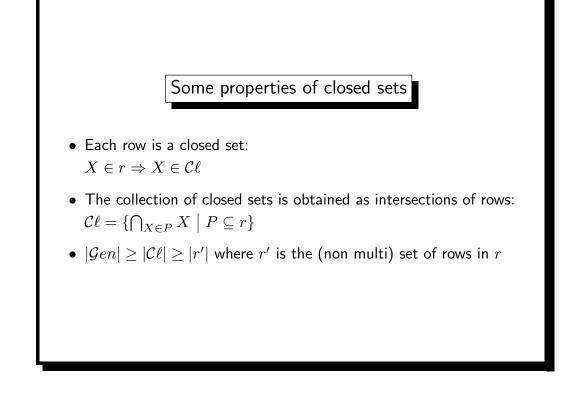
- generators (also called *key patterns*) are a complementary concept
- item set X is a generator of X^+ iff there is no proper subset $Y \subset X$ such that $Y^+ = X^+$
- the collection of all generators:

$$\mathcal{G}en = \{ X \subseteq R \mid X^+ \neq Y^+ \text{ for all } Y \subset X \}$$

- generators, too, are a sufficient representation for all sets:
- $fr(X) = \min\{fr(Y) \mid Y \in \mathcal{G}en \text{ and } Y \subseteq X\}$
- discovery of only frequent closed sets or frequent generators can be much more efficient than explicit discovery of all frequent sets



9



Data mining, Autumn 2002, Closed sets, generators, condensed representations

Discovery of all frequent generators

- Lemma If $X \in \mathcal{G}en$ then $Y \in \mathcal{G}en$ for all subsets $Y \subseteq X$
- thus: being a generator is a downwards monotone property, just like being a frequent set

 \Rightarrow the levelwise algorithm and Apriori in special are directly applicable

• recall Apriori algorithm:

1.
$$C_1 := \{\{A\} \mid A \in R\};$$

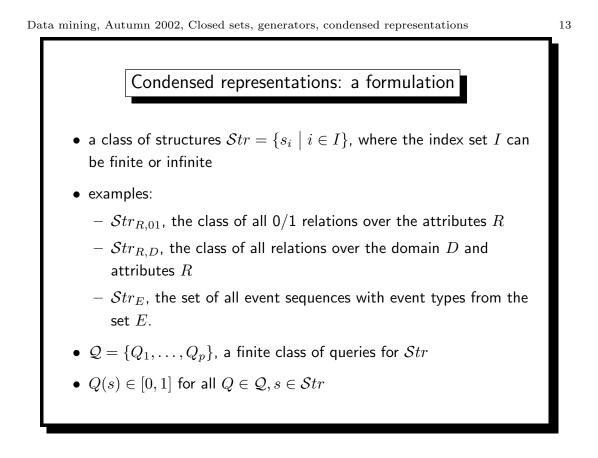
- 2. l := 1;
- 3. while $C_l \neq \emptyset$ do
- 4. compute $\mathcal{F}_l(r) := \{ X \in \mathcal{C}_l \mid fr(X, r) \geq min_fr \};$
- 5. l := l + 1;
- 6. compute $C_l := C(\mathcal{F}_{l-1}(r));$
- 7. for all l and for all $X \in \mathcal{F}_l(r)$ do output X and fr(X, r);
- refine $\mathcal{F}_l(r)$ and Step 4 to select frequent generators:
 - 4. compute $\mathcal{F}_l(r) := \{X \in \mathcal{C}_l \mid fr(X, r) \ge \min_f r \text{ and } fr(X, r) \neq fr(Y, r) \text{ for all } Y \subset X\};$
- add a step that outputs generators in the negative border:
 - for all l and for all X ∈ C_l \ F_l(r) such that fr(X,r) < min_fr do output "X is in Bd⁻(Gen ∩ F(r, min_fr))";

Data mining, Autumn 2002, Closed sets, generators, condensed representations

- the negative border is needed above to determine (border of) the collection of frequent sets:
- X is frequent iff there is no Y in the border such that $Y \subseteq X$
- otherwise the frequency of X is the minimum of the frequencies of its subsets in the output of the algorithm
- frequent generators and the negative border are a *condensed representation* of frequent sets

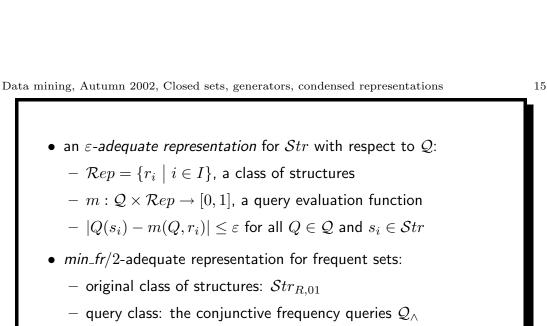
11

- the easy way:
 - 1. find all frequent generators
 - 2. compute closures of the generators from the database



• example query classes for $Str_{R,01}$:

- the *disjunctive* queries



- the conjunctive queries $\mathcal{Q}_{\wedge} = \{Q_X : r \mapsto fr(X, r) \mid X \subseteq R\}$

 $\mathcal{Q}_{\vee} = \{Q'_X : r \mapsto \frac{|\{t \in r \mid t[A] = 1 \text{ for some } A \in X\}|}{|r|} \mid X \subseteq R\}$

- condensed class of structures: frequent closed sets $\mathcal{F}(r, \min_{-}fr) \cap \mathcal{C}\ell$
- query evaluation function of $Q_X \in \mathcal{Q}_{\wedge}$: $r \mapsto \max(\{fr(Y,r) \mid Y \in \mathcal{F}(r, \textit{min_fr}) \cap \mathcal{C}\ell \text{ and } X \subseteq Y\} \cup \{\textit{min_fr}/2\})$

- Lossless (0-adequate) condensed representations for frequent sets and their frequencies
 - frequent generators and their negative border
 - frequent closed sets
- Lossless condensed representations for the collection of frequent sets (not frequencies)
 - positive border
 - negative border

Data mining, Autumn 2002, Closed sets, generators, condensed representations

- Approximate (ε -adequate, $\varepsilon \ge 0$) condensed representations:
 - a random sample
 - $-~\delta\text{-free}$ sets, almost closures
 - disjunction-free sets, disjunction-free generators
 - ...

17



- idea: relax the definition of closure
- B is "almost" in the closure of A if $|\mathcal{M}(\{A\},r)| |\mathcal{M}(\{A,B\},r)| \le \delta$
- $\bullet\,$ do not output X if it is almost in the closure of some other set
- allows limited approximation error; can reduce the size of output and running time considerably

		Expe	erimental res	sults		
		-				
Dataset, <i>min_fr</i>	$ \mathcal{F}(r, \min_{r}) $	db scans	$ \mathcal{F}(r, \textit{min}_fr) \cap \mathcal{C}\ell $	db scans	$ \mathcal{F}(r, \textit{min_fr}) \cap 4 - \mathcal{C}\ell $	db scans
ANPE, 0.005			412092	11	182829	10
ANPE, 0.05	25781	11	11125	9	10931	9
ANPE, 0.1	6370	10	2898	8		
	1516	9	638	7		
ANPE, 0.2	1510					
ANPE, 0.2 census, 0.005	1510		85950	9	39036	8
	90755	13	85950 10513	9 9	39036 5090	
census, 0.005		13 12				8



Literature

- Closed sets and generators:
 N. Pasquir et al.: Discovering frequent closed itemsets for association rules, ICDT 1999.
- δ-free sets/almost closures:
 J-F. Boulicaut et al.: Approximation of frequency queries by means of free-sets. PKDD 2000.
- (Condensed representations:
 H. Mannila and H. Toivonen: Multiple uses for frequent sets and condensed representations, KDD 1996.)
- (Original work on closed sets also by M. Zaki et al.)