Lec 07: Compressing Itemsets

COSC–254 – February 18, 2019

Outline

The itemsets explosion and what to do about it

Maximal Frequent Itemsets

Closed Frequent Itemsets

The pattern flood

Consider the following dataset

tid	A	B	C	D	E	F	G	H
1	✔	✔	✔	✔	✔			
2		✔	✔	✔	✔	✔	✔	
3			✔	✔	✔	✔	✔	✔
4	✔	✔			✔	✔	✔	✔
5		✔	✔		✔	✔		✔
6	✔			✔	✔	✔		✔
7	✔	✔	✔	✔	✔	✔	✔	✔

Image by J. Vreeken.

How many itemsets with support at least 1?
255

How many itemsets with freq. at least $1/2$?
31

"The goal of DM is to *summarize* data".

In these cases, the set of FIs is hardly a summary.

The wine explosion

The `wine` dataset has 178 transactions, built upon 14 items.

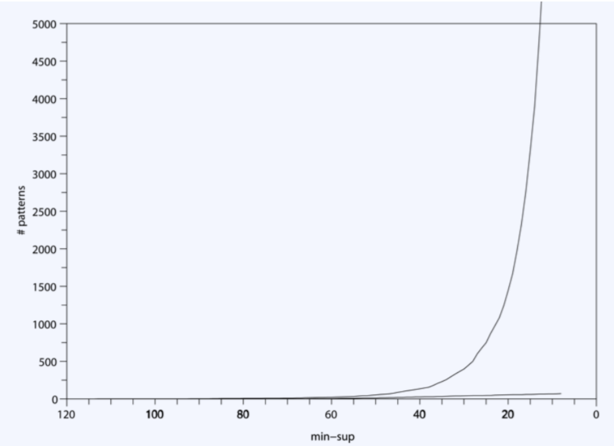


Image by J. Vreeken.

Pattern Explosion

With *high* min.supp. thresholds, you find *only few* patterns.

Most of them are "common knowledge" and not very interesting (e.g., {bread, milk })

With *low* min.supp. thresholds, you find an *very large* number of patterns.

Many are *potentially* interesting, but:

1) many are *noise*: happen too few times to *generalize* from them
2) there are *orders of magnitude* more patterns than rows in the data.

Curbing the explosion

IDEA: *compress* the collection of FIs

How:

1) Impose a strict *local criterion* for patterns, to remove *locally redundant* patterns.
2) Return only the non-redundant patterns

Depending on the local criterion, the compression may be *lossless* or *lossy*.

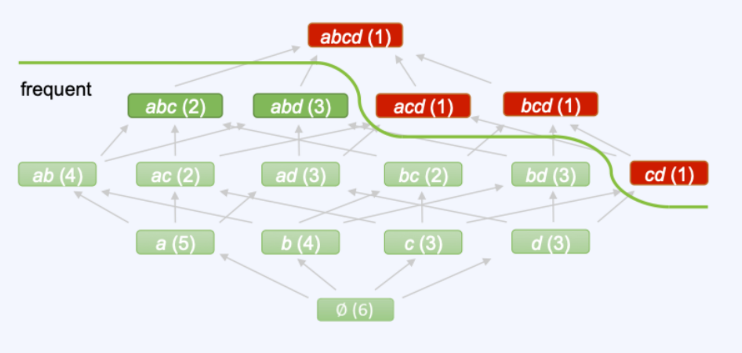
Outline

✔ The itemsets explosion and what to do about it

$$\left.\begin{array}{l}\text{Maximal Frequent Itemsets}\\[1em]\text{Closed Frequent Itemsets}\end{array}\right\}\quad\text{Different local criteria}$$

C.R.E.A.M.: Compression Rules Everything Around Me

Q: Is there a subset of $\mathsf{FI}(\mathcal{D}, \ell)$ from which we can *reconstruct* $\mathsf{FI}(\mathcal{D}, \ell)$?



A: the set of FIs for which *no superset* is a FI.

Image by J. Vreeken

Maximal Frequent Itemsets

Definition (Maximal Frequent Itemset)

A FI $A \in \mathsf{FI}(\mathcal{D}, \ell)$ is *maximal* if and only if none of its supersets is frequent. I.e.,

$$\text{for each } B \supset A, \text{ it holds } \mathsf{supp}_{\mathcal{D}}(B) < \ell$$

Obtaining the MFIs can be done easily by post-processing $\mathsf{FI}(\mathcal{D}, \ell)$.

Maximal Frequent FIs: Example

DEF: A FI is *maximal* if and only if none of its supersets is frequent.

Let the min. supp. thres. be 3. Which of the following itemsets are maximal and why?

Itemset	Support	Maximal?	Why?
$\{a\}$	4	✗	$\{a, b\}$ is frequent
$\{b\}$	5	✗	$\{a, b\}$ is frequent
$\{c\}$	3	✗	$\{b, c\}$ is frequent
$\{a, b\}$	4	✔	$\{a, b, c\}$ is only superset and is not frequent
$\{a, c\}$	2	✗	Not frequent
$\{b, c\}$	3	✔	$\{a, b, c\}$ is only superset and is not frequent
$\{a, b, c\}$	2	✗	Not frequent

Example from slides at `http://mmds.org`.

Reconstructing the FIs

With the set of MFIs we can reconstruct $\mathsf{FI}(\mathcal{D}, \ell)$. How?

$\mathsf{FI}(\mathcal{D}, \ell)$ contains all and only the *subsets of each MFI*:

X is frequent if and only if there exists a MFI Y such that $X \subseteq Y$.

Q: Do we lose any information?

Yes: no information about the *support* of frequent-but-not-maximal X.

MFIs are a *lossy* representation of $\mathsf{FI}(\mathcal{D}, \ell)$.

Outline

✔ The itemsets explosion and what to do about it

✔ Maximal Frequent Itemsets (lossy)

Closed Frequent Itemsets (*lossless*)

Local criteria

The "local" criterion for MFIs is not very local.

Lack of locality causes the loss in information.

SOLUTION: find a more local criterion.

We need to somehow be able to keep track of "where" in the lattice the support *changes*

IDEA: if A and B, with $B \supset A$, have the *same support,* then
A *and its support* can be reconstructed from B.

No loss of information!

Closed Frequent Itemsets

Definition (Closed Frequent Itemset)

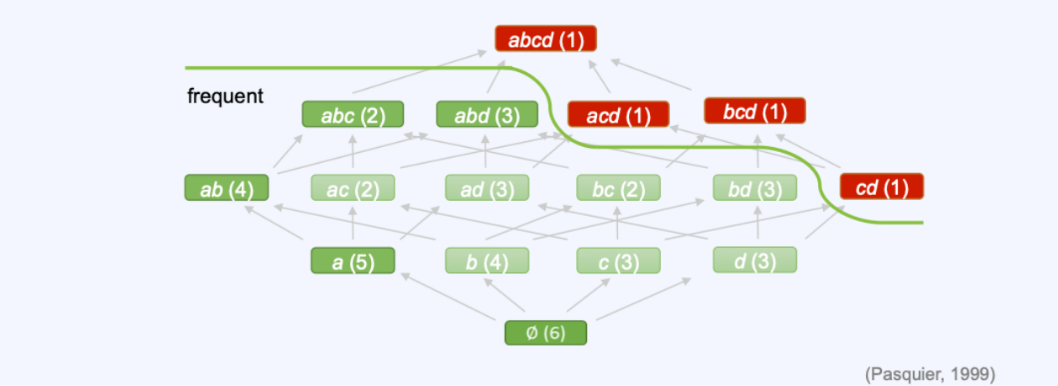
An itemset $A \in \mathsf{FI}(\mathcal{D}, \ell)$ is a *Closed Frequent Itemset* (CFI) iff *all* its supersets have *smaller* support than A. I.e.,

$$\text{for each } B \supset A, \mathsf{supp}_{\mathcal{D}}(B) < \mathsf{supp}_{\mathcal{D}}(A)$$

Finding the set of CFIs can be done by post-processing of $\mathsf{FI}(\mathcal{D}, \ell)$.

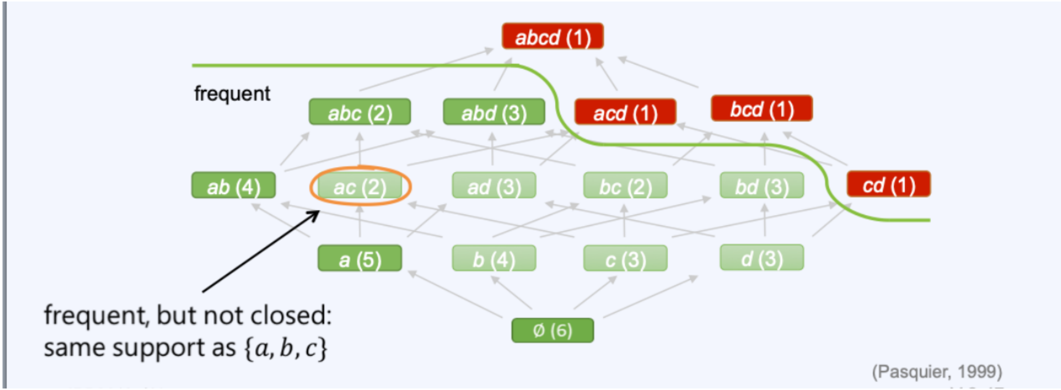
Quite expensive if done naïvely, but there are smart algorithms (Charm)

Closed Frequent Itemsets: Example



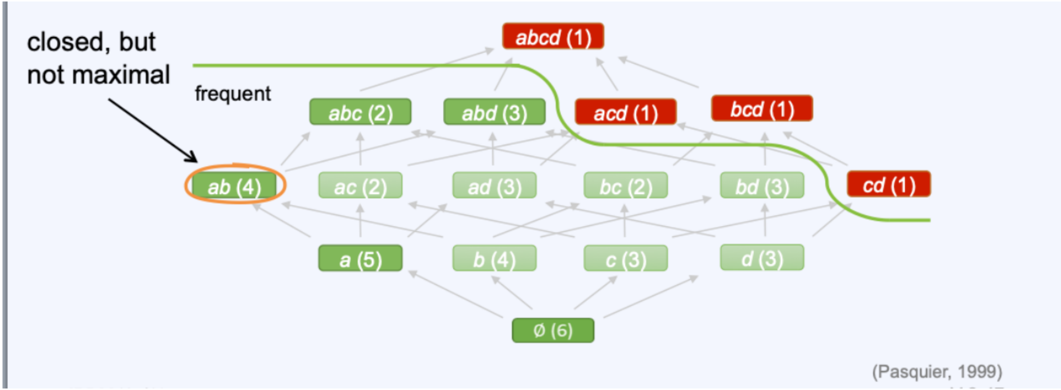
Images by J. Vreeken

Closed Frequent Itemsets: Example



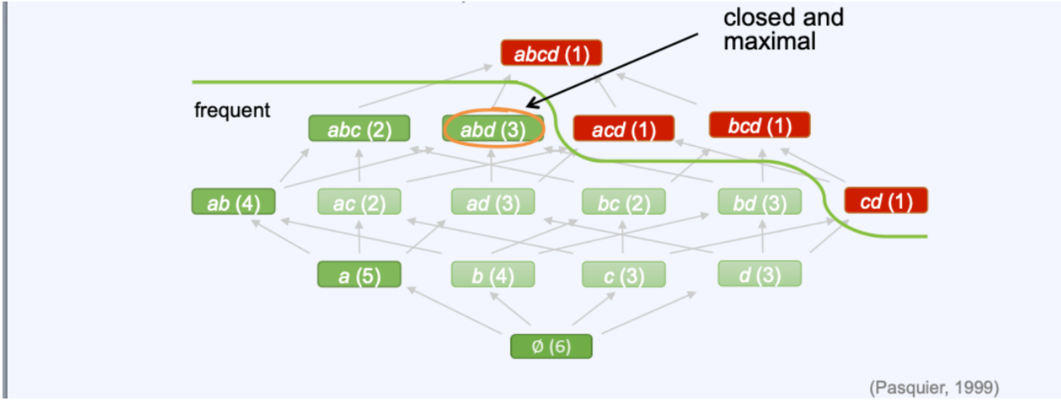
Images by J. Vreeken

Closed Frequent Itemsets: Example



Images by J. Vreeken

Closed Frequent Itemsets: Example



Images by J. Vreeken

Reconstructing the FIs

Given *all* closed frequent itemsets we can reconstruct $\mathsf{FI}(\mathcal{D}, \ell)$ *including the supports*:

- X is frequent if it is a subset of a closed frequent itemset;
- $\mathsf{supp}_{\mathcal{D}}(X) = \max\{\mathsf{supp}_{\mathcal{D}}(Z) \;:\; X \subseteq Z, Z \text{ is frequent and closed}\}$

Why “closed”?

Consider the following functions:

$t(X)$ returns all transactions that contain itemset X

$i(T)$ returns all *items* that are contained in *every* transaction of a set T of transactions

The *closure function* $c(X)$ maps itemsets to itemsets by

$$c(X) = (i \circ t)(X) = i(t(X))$$

The closure function is:

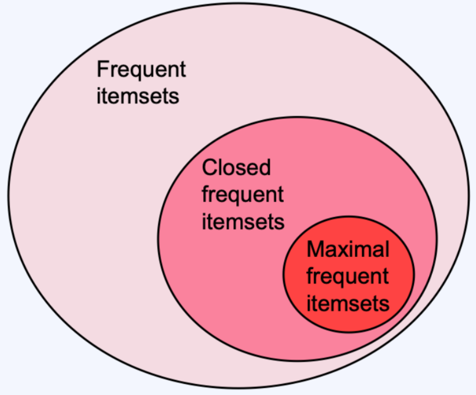
extensive: $X \subseteq c(X)$

monotonic: if $X \subseteq Y$, then $c(X) \subseteq c(Y)$

idempotent: $c(c(X)) = c(X)$.

Itemset X is closed if and only if $X = c(X)$.

Itemset Taxonomy



There are many other ways of summarizing the FIs:

Non-Derivable FIs, constrained FIs, itemsets that compress, ...

Image by J. Vreeken

Outline

- ✔ The itemsets explosion and what to do about it
- ✔ Maximal Frequent Itemsets (lossy representation of $\mathsf{FI}(\mathcal{D}, \ell)$)
- ✔ Closed Frequent Itemsets (*lossless* representation)

Pattern mining recap

- Itemsets: basic definitions, support,
- Anti-monotonicity of the support
- Apriori and Eclat
- Association rules: basic definitions, support, confidence
- Pruning of rules by support
- Pruning of rules by confidence
- Algorithm to mine ARs
- Pattern explosion
- Maximal Frequent Itemsets
- Closed Frequent Itemsets