# Chapter 4: Frequent Itemsets and Association Rules

Jilles Vreeken

Revision 1, November 9th
Notation clarified, Chi-square: clarified

Revision 2, November 10th
details added of derivability example

Revision 3, November 12th
typo fixed in Pearson correlation

IRDM '15/16

5 Nov 2015

UNIVERSITÄT DES SAARLANDES

M2CI CLUSTER OF EXCELLENCE

mpi max planck institut informatik

# Recall the Question of the week

How can we mine **interesting patterns** and **useful rules** from data?

# IRDM Chapter 4, today

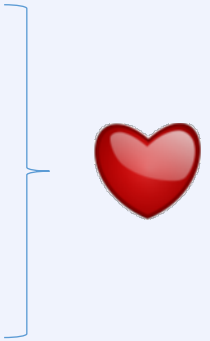
1. Definitions
2. Algorithms for Frequent Itemset Mining
3. Association Rules and Interestingness
4. Summarising Collections of Itemsets

You'll find this covered in
Aggarwal Chapter 4, 5.2
Zaki & Meira, Ch. 10, 11

# Chapter 4.3:
# Association Rules

# IRDM Chapter 4.3

1. Generating Association Rules
2. Measures of Interestingness
3. Properties of Measures
4. Simpson's Paradox



You'll find this covered in
Aggarwal, Chapter 4
Zaki & Meira, Ch. 10

# Generating Association Rules

We can generate association rules from frequent itemsets

- if $Z$ is a frequent itemset and $X \subset Z$ is its proper subset, we have rule $X \rightarrow Y$, where $Y = Z \setminus X$

These rules are frequent because

$$supp(X \rightarrow Y) = supp(X \cup Y) = supp(Z)$$

- we still need to compute the confidence as $\frac{supp(Z)}{supp(X)}$

Which means, if rule $X \rightarrow Z \setminus X$ is not confident, no rule of type $W \rightarrow Z \setminus W$, with $W \subseteq X$, is confident

- we can use this to prune the search space

# Pseudo-code

**ALGORITHM 8.6. Algorithm ASSOCIATIONRULES**

```
ASSOCIATIONRULES (F, minconf):
1  foreach Z ∈ F, such that |Z| ≥ 2 do
2      A ← {X | X ⊂ Z, X ≠ ∅}
3      while A ≠ ∅ do
4          X ← maximal element in A
5          A ← A \ X // remove X from A
6          c ← sup(Z)/sup(X)
7          if c ≥ minconf then
8              print X ⟶ Y, sup(Z), c
9          else
10             A ← A \ {W | W ⊂ X} // remove all subsets of X from A
```

(Algorithm 8.6 in Zaki & Meira)

# Measures of interestingness

Consider the following example:

| | Coffee | Not Coffee | $\sum$ |
|---|---|---|---|
| **Tea** | 150 | 50 | **200** |
| **Not Tea** | 650 | 150 | **800** |
| $\sum$ | **800** | **200** | **1000** |

Rule $\{\text{Tea}\} \rightarrow \{\text{Coffee}\}$ has 15% support and 75% confidence

- reasonably good numbers

Is this a good rule?

- the overall fraction of coffee drinkers is 80%,
  drinking tea **reduces** the probability of drinking coffee!

# Problems with confidence

The support-confidence framework does not take the support of the consequent into account

- rules with relatively small support for the antecedent and high support for the consequent often have high confidence

To fix this, many other measures have been proposed

Most measures are easy to express using **contingency tables**

We'll use $s_{ij}$ as shorthand for support:
$s_{11} = supp(AB)$, $s_{01} = supp(\neg AB)$, …

Analogue, we'll say $f_{ij}$ for frequency:
$f_{11} = freq(AB), f_{01} = freq(\neg AB)$, …

| | **B** | **¬B** | **∑** |
|---|---|---|---|
| **A** | $s_{11}$ | $s_{10}$ | $s_{1+}$ |
| **¬A** | $s_{01}$ | $s_{00}$ | $s_{0+}$ |
| **∑** | $s_{+1}$ | $s_{+0}$ | $N$ |

(revised on Nov 9th, now using $s_{ij}$ notation to more clearly indicate **support**)

# Statistical Coefficient of Correlation

A natural statistical measure between a pair of items is the **Pearson correlation coefficient**

$$\rho = \frac{E[XY] - E[X]E[Y]}{\sigma(X)\sigma(Y)}$$

$$= \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - E[X]^2}\sqrt{E[Y^2] - E[Y]^2}}$$

# Pearson of Correlation of Items

For items $A$ and $B$ it reduces to

$$\rho_{AB} = \frac{f_{11} - f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}(1 - f_{1+})(1 - f_{+1})}}$$

It is $+1$ when the data is perfectly positively correlated, -1 when perfectly negatively correlated, and 0 when uncorrelated.

(revised on November 12th; typo fixed, as $f_{11}$ should be inside the nominator)

# Chi-square

$\mathcal{X}^2$ is another natural statistical measure of significance for itemsets. For a set of $k$ items, it compares the observed frequencies against the expected frequencies of all $2^k$ possible states.

$$\mathcal{X}^2(X) = \sum_{Y \in \mathcal{P}(X)} \frac{(freq(Y) - E_X[freq(Y)])^2}{E_X[freq(Y)]}$$

where $\mathcal{P}(X)$ is the powerset of $X$ and $E_X[freq(Y)]$ is the expected frequency of state $Y$ over itemset $X$

For example, for $X = \{\text{beer}, \text{diapers}\}$, it considers states $\{\text{beer}, \text{diapers}\}$, $\{\neg\text{beer}, \text{diapers}\}$, $\{\text{beer}, \neg\text{diapers}\}$ and $\{\neg\text{beer}, \neg\text{diapers}\}$.

(Brin et al. 1998, 1.6k+ cites)

(revised on Nov 9th, now using $E_X[freq(Y)]$ to more clearly indicate the expectation is of state $Y$ over itemset $X$)

# Chi-square (2)

To compute $\mathcal{X}^2(X)$ we need to define $E_X[freq(Y)]$.

The standard way is to **assume independence** between the items of $Y$. That is, the probability of a state $Y$ is the multiplication of its individual item frequencies.

$$E_X[freq(Y)] = \prod_{A \in Y} freq(A) \prod_{A \in X \setminus Y} (1 - freq(A))$$

The first product is over the items that **are** present in $Y$ (the 1s). For these their empirical probability is simply $freq(\cdot)$.

The second product considers the 0s in $Y$, or in other words, the 1s of $X$ not in $Y$. The empirical probability of not seeing an item A is $(1 - freq(A))$.

Note! Independence between items is a very strong assumption, and hence we will find that many itemsets will be 'significantly' correlated.

(revised on Nov 9th, now using $E_X[freq(Y)]$ notation, added more explanation)

# Chi-square (3)

$$\mathcal{X}^2(X) = \sum_{Y \in \mathcal{P}(X)} \frac{(freq(Y) - E_X[freq(Y)])^2}{E_X[freq(Y)]}$$

Chi-square scores close to 0 indicate statistical independence, while larger values indicate stronger dependencies.

- no differentiation to positive or negative correlation
- it is **computationally costly** at $O(2^{|X|})$
- but as it is upward closed, we can mine interesting sets efficiently

Always be thoughtful of how you define your expected frequency!

(revised on Nov 9th, now using $E_X[freq(Y)]$ notation, added more explanation)

# Interest Ratio

The **interest ratio** $I$ of rule $A \rightarrow B$ is

$$I(A,B) = \frac{N \times supp(AB)}{supp(A) \times supp(B)} = \frac{Ns_{11}}{s_{1+}s_{+1}}$$

- it is equivalent to **lift** = $\frac{conf(A \rightarrow B)}{supp(B)}$

Interest ratio compares the frequencies against the assumption that $A$ and $B$ are independent

- if $A$ and $B$ are independent, $s_{11} = \frac{s_{1+}s_{+1}}{N}$

Interpreting interest ratios

- $I(A,B) = 1$ if $A$ and $B$ are independent
- $I(A,B) > 1$ if $A$ and $B$ are positively correlated
- $I(A,B) < 1$ if $A$ and $B$ are negatively correlated

($f_{ij}$ changed into $s_{ij}$ in revision 1)

# The cosine measure

The **cosine**, or $IS$, measure of rule $A \rightarrow B$ is defined as

$$cosine(A,B) = \sqrt{I(A,B) \times supp(AB)/N} = \frac{s_{11}}{\sqrt{s_{1+}} \times \sqrt{s_{+1}}}$$

which is regular cosine if we think of $A$ and $B$ as binary vectors

It also is the **geometric mean** between the confidences of $A \rightarrow B$ and $B \rightarrow A$ as

$$cosine(A,B) = \sqrt{\frac{supp(AB)}{supp(A)} \times \frac{supp(AB)}{supp(B)}} = \sqrt{conf(A \rightarrow B) \times conf(B \rightarrow A)}$$

($f_{ij}$ changed into $s_{ij}$ in revision 1)

# Examples (1)

| | Coffee | Not Coffee | $\sum$ |
|---|---|---|---|
| **Tea** | 150 | 50 | **200** |
| **Not Tea** | 650 | 150 | **800** |
| $\sum$ | **800** | **200** | **1000** |

The interest ratio of $\{\text{Tea}\} \rightarrow \{\text{Coffee}\}$ is $\frac{1000 \times 150}{800 \times 200} = 0.9375$

- almost 1, so not very interesting;
  below 1, so (slight) negative correlation

The $cosine$ of this rule, however, is $0.375$

- quite far from 0, so, it **is** interesting.

# Examples (2)

| | *p* | ¬*p* | Σ |
|---|---|---|---|
| *q* | 880 | 50 | 930 |
| ¬*q* | 50 | 20 | 70 |
| Σ | 930 | 70 | 1000 |

| | *r* | ¬*r* | Σ |
|---|---|---|---|
| *t* | 20 | 50 | 70 |
| ¬*t* | 50 | 880 | 930 |
| Σ | 70 | 930 | 1000 |

$I(p,q) = 1.02$ and $I(r,t) = 4.08$

- $p$ and $q$ are close to independent
- $r$ and $t$ have highest interest factor

But $p$ and $q$ appear together in 88% of cases
But $r$ and $t$ appear together only seldom

Now $conf(p \rightarrow q) = 0.946$ and $conf(r \rightarrow t) = 0.286$

(revised on Nov 9th, now using $t$ instead of $s$ to avoid confusion with support-notation)

# Examples (2)

| | *p* | ¬*p* | Σ |
|---|---|---|---|
| *q* | 880 | | |
| ¬*q* | 5 | | |
| Σ | 9 | | |

| | *r* | ¬*r* | Σ |
|---|---|---|---|
| | | | 70 |
| | | | 930 |
| | | | 1000 |

Bottom line: **Lunch is not free.** There is no single measure that works well all the time.

$I(p, q)$

- $p$ and
- $r$ and

ar
f cases
ear
together only seldom

Now $conf(p \rightarrow q) = 0.946$ and $conf(r \rightarrow s) = 0.286$

# Measures for pairs of itemsets

| Measure (symbol) | Definition |
|---|---|
| Correlation ($\phi$) | $\frac{Ns_{11} - s_{1+}s_{+1}}{\sqrt{s_{1+}s_{+1}s_{0+}s_{+0}}}$ |
| Odds ratio ($\alpha$) | $(s_{11}s_{00})/(s_{10}s_{01})$ |
| Kappa ($\kappa$) | $\frac{Ns_{11} + Ns_{00} - s_{1+}s_{+1} - s_{0+}s_{+0}}{N^2 - s_{1+}s_{+1} - s_{0+}s_{+0}}$ |
| Interest ($I$) | $(Ns_{11})/(s_{1+}s_{+1})$ |
| Cosine ($cosine$) | $(s_{11})/(\sqrt{s_{1+}s_{+1}})$ |
| Pieatetsk-Shapiro ($PS$) | $\frac{s_{11}}{N} - \frac{s_{1+}s_{+1}}{N^2}$ |
| Collective Strength ($S$) | $\frac{s_{11} + s_{00}}{s_{1+}s_{+1} + s_{0+}s_{+0}} \times \frac{N - s_{1+}s_{+1} - s_{0+}s_{+0}}{N - s_{11} - s_{00}}$ |
| Jaccard ($J$) | $s_{11}/(s_{1+} + s_{+1} - s_{11})$ |
| All-confidence ($h$) | $\min\left[\frac{s_{11}}{s_{1+}}, \frac{s_{11}}{s_{+1}}\right]$ |

(revised on Nov 9th, now using $s_{ij}$ notation to more clearly indicate **support**) (after Tan, Steinbach, Kumar, Table 6.12)

# Measures for association rules

| Measure (symbol) | Definition |
|---|---|
| Goodman-Kruskal ($\lambda$) | $\left(\sum_j \max_k s_{jk} - \max_k s_{+k}\right) / (N - \max_k s_{+k}$ |
| Mutual Information ($M$) | $\left(\sum_i \sum_j \frac{s_{ij}}{N} \log \frac{N s_{ij}}{s_{i+} s_{+j}}\right) / \left(-\sum_i \frac{s_{i+}}{N} \log \frac{s_{i+}}{N}\right)$ |
| J-Measure ($J$) | $\frac{s_{11}}{N} \log \frac{N s_{11}}{s_{1+} s_{+1}} + \frac{s_{10}}{N} \log \frac{N s_{10}}{s_{1+} s_{+0}}$ |
| Gini index ($G$) | $\frac{s_{1+}}{N} \times \left[\left(\frac{s_{11}}{s_{1+}}\right)^2 + \left(\frac{s_{10}}{s_{1+}}\right)^2\right] - \left(\frac{s_{+1}}{N}\right)^2 + \frac{s_{0+}}{N} \times \left[\left(\frac{s_{01}}{s_{0+}}\right)^2 + \left(\frac{s_{00}}{s_{0+}}\right)^2\right] - \left(\frac{s_{+0}}{N}\right)^2$ |
| Laplace ($L$) | $(s_{11} + 1)/(s_{1+} + 2)$ |
| Conviction ($V$) | $(s_{1+} s_{+0})/(N s_{10})$ |
| Certainty factor ($F$) | $\left(\frac{s_{11}}{s_{1+}} - \frac{s_{+1}}{N}\right) / \left(1 - \frac{s_{+1}}{N}\right)$ |
| Added Value ($AV$) | $\frac{s_{11}}{s_{1+}} - \frac{s_{+1}}{N}$ |

(revised on Nov 9th, now using $s_{ij}$ notation to more clearly indicate **support**) (after Tan, Steinbach, Kumar, Table 6.12)

# Properties of Measures

Most measures do not agree on how they rank itemset pairs or rules

To understand how they behave, we need to study their properties

- measures that share some properties behave similarly under that property's conditions

# Three properties

A measure has the **inversion property** if its value stays the same if we exchange $s_{11}$ with $s_{00}$ and $s_{01}$ with $s_{10}$

- the measure is invariant for flipping bits – it is **bit symmetric**

A measure has the **null addition property** if is not affected by increasing $s_{00}$ if other values stay constant

- the measure is invariant on adding new transactions that have an empty intersection with the itemset

A measure has the **scaling invariance property** if it is not affected by replacing the values $s_{11}$, $s_{10}$, $s_{01}$ and $s_{00}$ with values $k_1k_3s_{11}$, $k_2k_3s_{10}$, $k_1k_4s_{01}$, and $k_2k_4s_{00}$

- where all $k_i$ are positive constants

(revised on Nov 9th, now using $s_{ij}$ notation to more clearly indicate **support**)

# Which properties hold?

| Symbol | Measure | Inversion | Null Addition | Scaling |
|---|---|---|---|---|
| $\phi$ | $\phi$-coefficient | Yes | No | No |
| $\alpha$ | odds ratio | Yes | No | Yes |
| $\kappa$ | Cohen's | Yes | No | No |
| $I$ | Interest | No | No | No |
| $IS$ | Cosine | No | Yes | No |
| $PS$ | Piatetsky-Shapiro's | Yes | No | No |
| $S$ | Collective strength | Yes | No | No |
| $\zeta$ | Jaccard | No | Yes | No |
| $h$ | All-confidence | No | No | No |
| $s$ | Support | No | No | No |

(Tan, Steinbach, Kumar, Table 6.17)

# Simpson's Paradox

Consider this data on sales of HDTVs and exercise machines

| | Exercise Machine | No Exercise Machine | $\sum$ |
|---|---|---|---|
| **HDTV** | 99 | 81 | **180** |
| **No HDTV** | 54 | 66 | **120** |
| $\sum$ | **153** | **147** | **300** |

$\{\text{HDTV}\} \rightarrow \{\text{Exerc. mach.}\}$ has confidence 0.55

$\{\neg\text{HDTV}\} \rightarrow \{\text{Exerc. mach.}\}$ has confidence 0.45

**Customers who buy HDTVs are more likely to also buy an exercise machines than those who don't buy HDTVs**

# Deeper Analysis

| | | Exerc. mach. | | |
|---|---|---|---|---|
| **Group** | **HDTV** | **Yes** | **No** | $\sum$ |
| **College** | **Yes** | 1 | 9 | 10 |
| | **No** | 4 | 30 | 34 |
| **Working** | **Yes** | 98 | 72 | 170 |
| | **No** | 50 | 36 | 86 |

For college students

- $conf(HDTV \rightarrow Exerc.mach.) = 0.10$
- $conf(-HDTV \rightarrow Exerc.mach.) = 0.118$

For working adults

- $conf(HDTV \rightarrow Exerc.mach.) = 0.577$
- $conf(-HDTV \rightarrow Exerc.mach.) = 0.581$

HDTV is **not** made more likely by exercise machine!

# The paradox, and why it happens

In the combined data, HDTVs and exercise machines correlate **positively**. In the stratified data, they correlate **negatively**.

- this is Simpson's paradox

The explanation

- most customers were working adults
  - they also bought most HDTVs and exercise machines
- in the combined data this increased the correlation between HDTVs and exercise machines

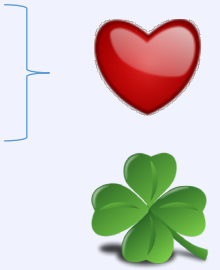
Moral of the story:

**Stratify your data properly!**

# Chapter 4.4: Summarising Collections of Itemsets

# IRDM Chapter 4.4

1. The Pattern Explosion
2. Maximal and closed frequent itemsets
3. Non-derivable frequent itemsets



You'll find this covered in
Aggarwal, Chapter 5.2
Zaki & Meira, Ch. 11 (non-derivable only here)

# The Pattern Flood

Consider the following table:

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

How many itemsets with minimum frequency of 1/7?

- 255 (!)

How many with minimum frequency of 1/2?

- 31 (!)

"The goal of data mining is … to summarize the data"

- Hardly a summary!

# The Pattern Explosion

This phenomenon is called **the pattern explosion**

For high thresholds you find only few patterns

- that only describe common knowledge

For lower thresholds you find enormously many patterns

- all potentially interesting
- many represent noise, and many will be highly redundant
- orders of magnitude more patterns than there are rows in the data

# Curbing the Explosion

There exist two main approaches

- frequent pattern summarisation

  - summarise the complete set of frequent patterns
  - impose a stricter **local criterion** for individual patterns that removes locally redundant patterns, e.g. closed frequent, maximal frequent
  - mine **all patterns** that satisfy this criterion

- pattern set mining

  - improves by imposing a **global criterion** for the complete result, e.g. shortest description of the data, minimal overlap, maximal entropy
  - mine that **set of patterns** that is optimal with regard to this criterion
  - this way we can **globally control** noise and redundancy

# Maximally frequent itemsets

Let $\mathcal{F}$ be the collection of all frequent itemsets for data $\boldsymbol{D}$

Itemset $X \in \mathcal{F}$ is **maximal** if it has no frequent supersets

- i.e. for all $Y \supset X$, $freq(Y) < minfreq$

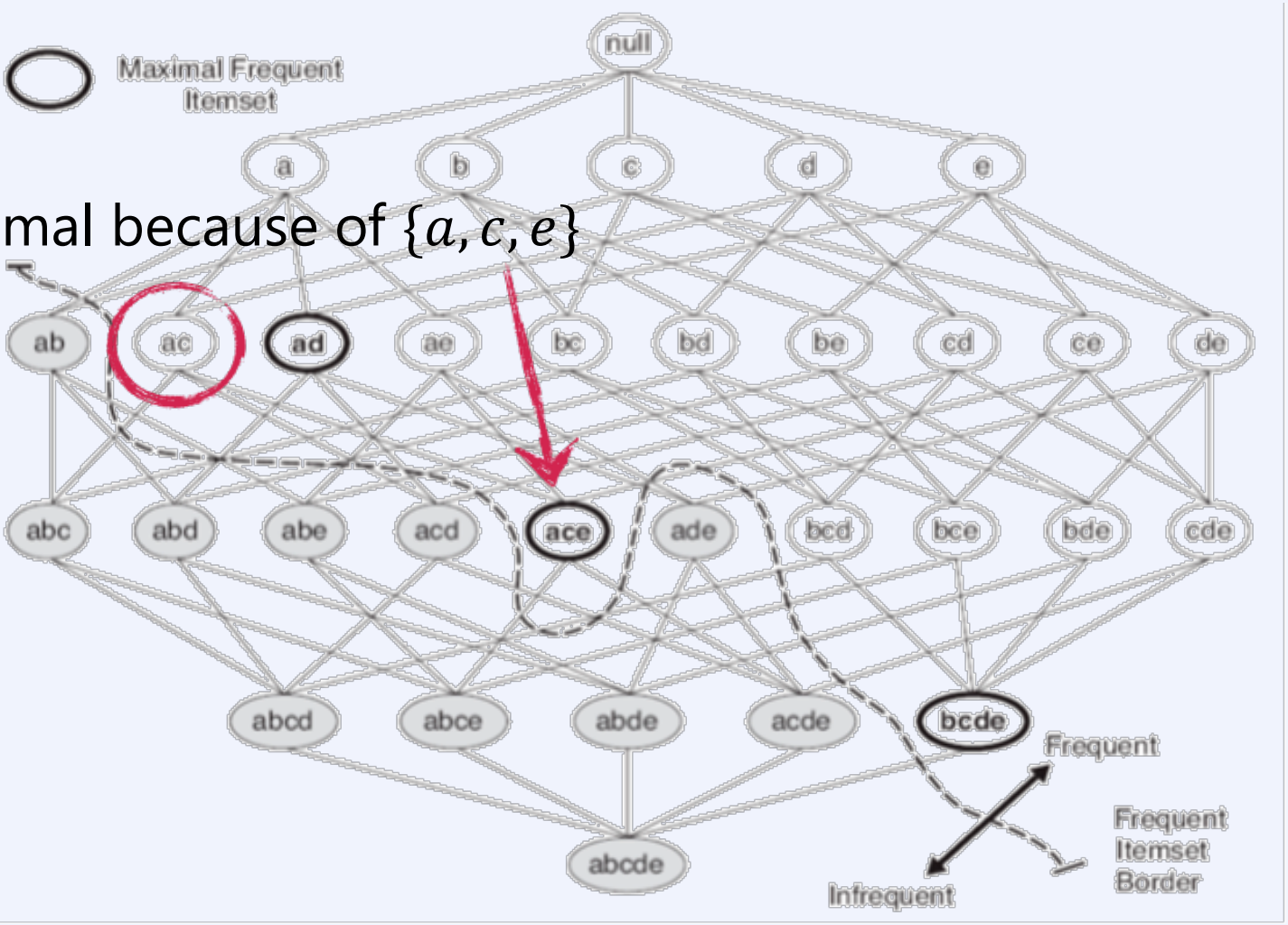
With the set of all maximal frequent itemsets
we can reconstruct all elements of $\mathcal{F}$

- $X$ is frequent if and only if there exists a maximal frequent itemset $M$ such that $X \subseteq M$
- this is a lossy representation:
  it does not tell us what the frequency of $X$ is

(Bayardo, 1998, 1.7k cites)

# Example of maximal frequent itemsets

Not maximal because of $\{a, c, e\}$



# Closed frequent itemsets

Let $\mathcal{F}$ be the collection of all frequent itemsets for data $\boldsymbol{D}$

Itemset $X \in \mathcal{F}$ is **closed** is all its supersets are less frequent

- i.e. for all $Y \supset X, freq(Y) < freq(X)$
- all maximal itemsets are also closed itemsets

Given the set of all frequent closed itemsets, we can reconstruct all elements of $\mathcal{F}$ including their frequency

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

(Pasquier et al. 1999, 1.5k cites)

# Why "closed"?

Consider the following functions

- $t(X)$ returns all transactions that contain itemset $X$
- $i(T)$ returns all items that are contained in all transactions in $T$

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

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

The closure function satisfies the following properties

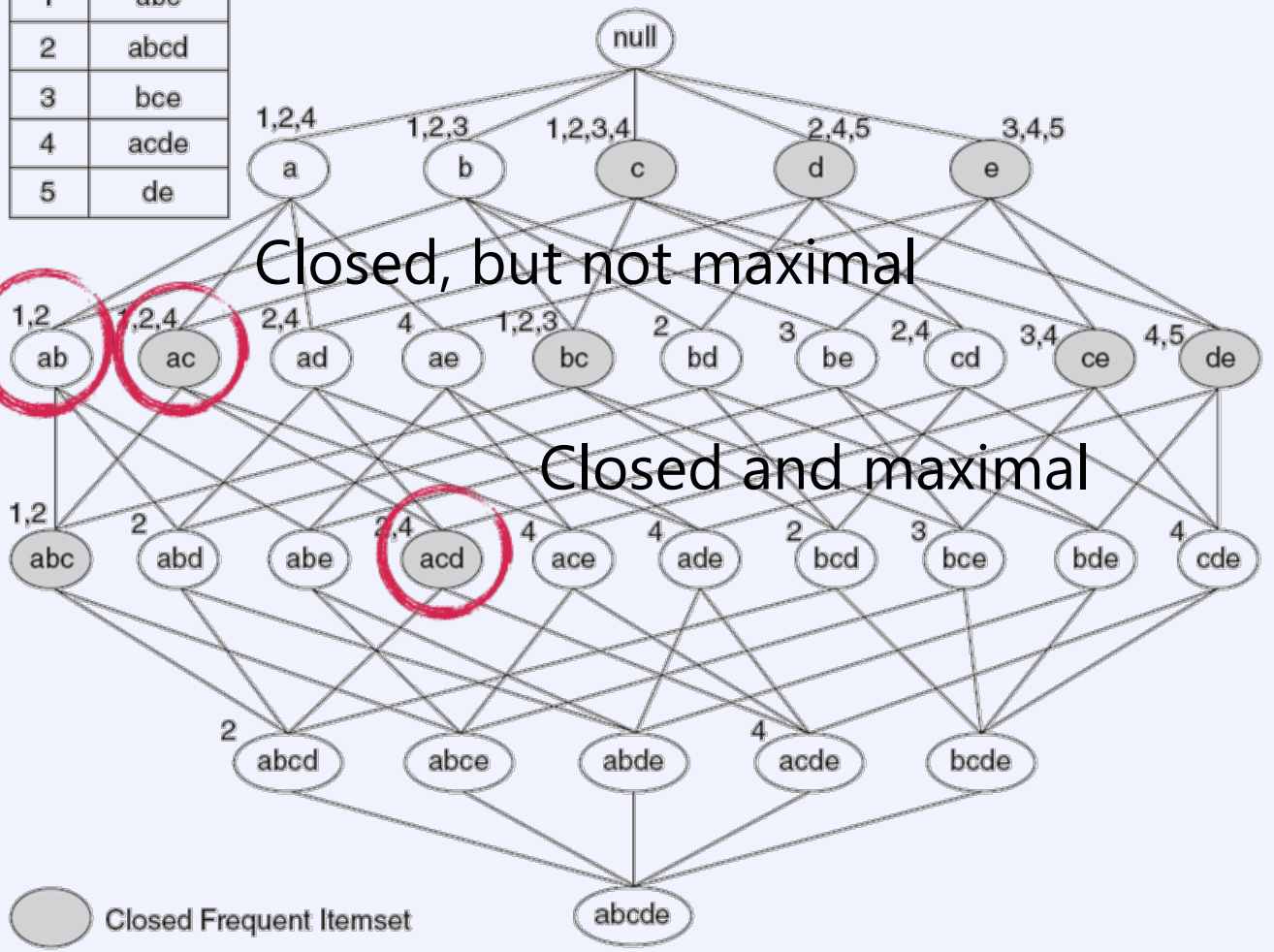
- extensive: $X \subset c(X)$
- monotonic: if $X \subseteq Y$, then $c(X) \subseteq c(Y)$
- Idempotent: $c\big(c(X)\big) = c(X)$

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

# Example of closed frequent itemsets

| TID | Items |
|---|---|
| 1 | abc |
| 2 | abcd |
| 3 | bce |
| 4 | acde |
| 5 | de |

minsup = 40%

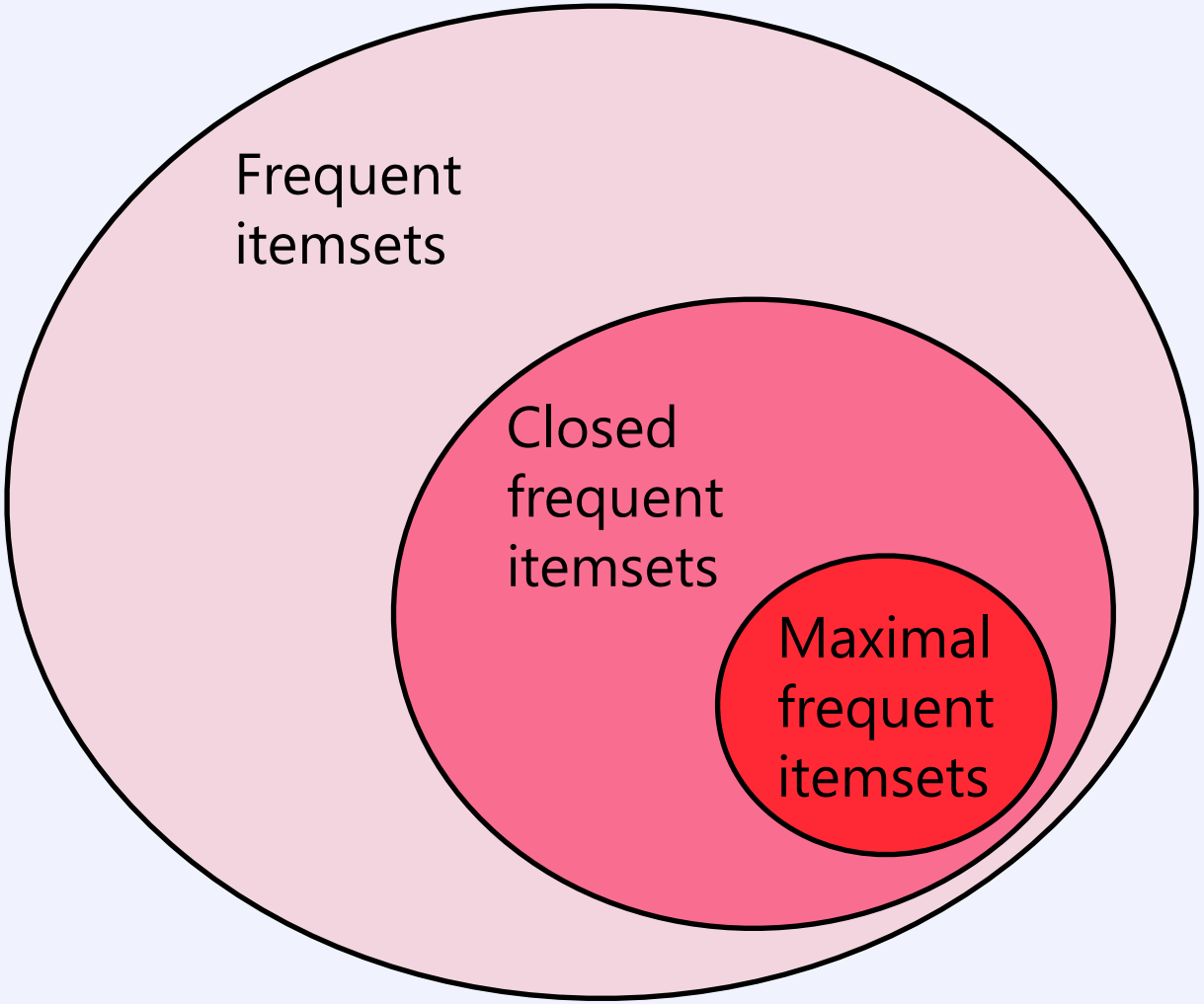



Closed, but not maximal

Closed and maximal

Itemset {a, b} is contained in transactions 1 and 2

# Itemset taxonomy



# Mining maximal and closed itemsets

Frequent maximal and closed itemsets can be found by post-processing the set of frequent itemsets

To find maximal itemsets:

- start with an empty set of candidate maximal itemsets $\mathcal{M}$
- **for each** frequent itemset $X \in \mathcal{F}$
  - **if** a superset of $X$ is in $\mathcal{M}$ , **continue**
  - **else** insert $X$ in $\mathcal{M}$ and remove all subsets of $X$ from $\mathcal{M}$
- **return** set $\mathcal{M}$

# Mining maximal and closed itemsets

Closed itemsets can be found from the frequent itemsets by computing their closure

- this can be very time consuming

The **Charm** algorithm avoids testing all frequent itemsets by using the following properties

- if $t(X) = t(Y)$, then $c(X) = c(Y) = c(X \cup Y)$
  - we can replace $X$ with $X \cup Y$ and prune $Y$
- if $t(X) \subset t(Y)$, then $c(X) \neq c(Y)$, but $c(X) = c(X \cup Y)$
  - we can replace $X$ with $X \cup Y$, but not prune $Y$
- if $t(X) \neq t(Y)$, $c(X) \neq c(Y) \neq c(X \cup Y)$
  - we cannot prune anything

(Zaki et al, 1999, 194 cites)

# Non-derivable frequent itemsets

Let $\mathcal{F}$ be the set of all frequent itemsets.

Itemset $X \in \mathcal{F}$ is **non-derivable** if we cannot derive its support from its subsets

- we can derive the support of $X$ if, by knowing the supports of all of the subsets of $X$, we can compute the support of $X$

If $X$ is derivable, it does not add any new information

- knowing just the non-derivable frequent itemsets, we can reconstruct every frequent itemset, including its frequency
- we only return itemsets that add new information on top of what we already knew

(Calders & Goethals, 2004, 121 citations)

# Support of a generalised itemset

A **generalised** itemset is an itemset of form $X\bar{Y}$

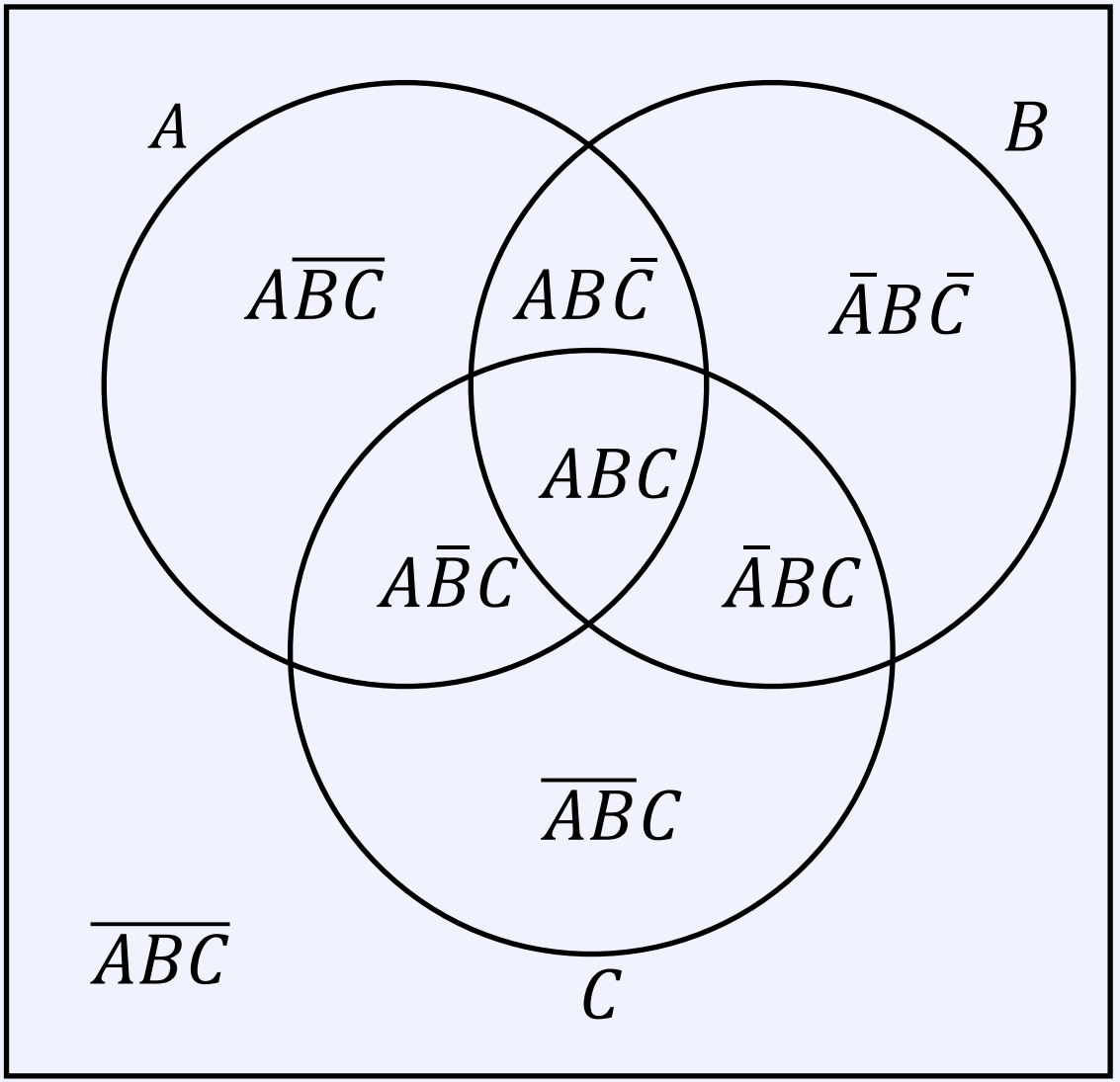
- all items in $X$ and none of the items in $Y$

The **support** of a generalised itemset $X\bar{Y}$ is the number of transactions that contain all items in $X$ but no items in $Y$

To compute the support of a generalised itemset $A\overline{BC}$ we

- take the support of $A$
- remove the supports of $AB$ and $AC$
- add the support of $ABC$ that was removed twice
- $supp(A\overline{BC}) = supp(A) - supp(AB) - supp(AC) + supp(ABC)$

# Generalised Itemsets



# The Inclusion-Exclusion Principle

Let $X\bar{Y}$ be a generalised itemset and $I = X \cup \mathrm{Y}$

Now, $supp(X\bar{Y})$ can be expressed as a combination of supports of supersets $J \supseteq X$ such that $J \subseteq I$ using the **inclusion-exclusion principle**

$$supp(X\bar{Y}) = \sum_{X \subseteq J \subseteq I} (-1)^{|J \setminus X|} supp(J)$$

For example, $supp(\overline{ABC}) = supp(\emptyset) - supp(A) - supp(B) - supp(C) + supp(AB) + supp(AC) + supp(BC) - supp(ABC)$

# Support Bounds

The inclusion-exclusion formula gives us bounds for the support of itemsets in $X \cup Y$ that are supersets of $X$

- all supports are non-negative!
- $supp(ABC) = supp(A) - supp(AB) - supp(AC) + supp(ABC) \geq 0$ implies $supp(ABC) \geq -supp(A) + supp(AB) + supp(AC)$
  - this is a lower bound, but we can also get upper bounds

In general, the bounds for itemset $I$ w.r.t. $X \subseteq I$

- if $|I \setminus X|$ is odd: $supp(I) \leq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} supp(J)$
- if $|I \setminus X|$ is even: $supp(I) \geq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} supp(J)$

# Deriving the Support

Given the formula for the bounds, we can define

- the **least upper bound** $lub(I)$ and
- the **greatest lower bound** $glb(I)$ for itemset $I$

We know that $supp(I) \in [glb(I), lub(I)]$

If $glb(I) = lub(I)$, then we can compute $supp(I)$ just knowing the support of subsets of $I$

- we say $I$ is **derivable**
- otherwise, $I$ is **non-derivable**

# Example deriving support – blackboard

| tid | A | B | C | D | E |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 0 | 1 | 1 |
| **2** | 0 | 1 | 1 | 0 | 1 |
| **3** | 1 | 1 | 0 | 1 | 1 |
| **4** | 1 | 1 | 1 | 0 | 1 |
| **5** | 1 | 1 | 1 | 1 | 1 |
| **6** | 0 | 1 | 1 | 1 | 0 |

Question: is itemset $ABC$ derivable?

# Example deriving support – blackboard

$supp(ABC) \geq 0$
$\geq s_{AB} + s_{AC} - s_A = 4 + 2 - 4 = 2$
$\geq s_{AC} + s_{BC} - s_B = 2 + 4 - 4 = 2$
$\geq s_{AB} + s_{BC} - s_B = 4 + 4 - 6 = 2$

$lb(ABC) = \{2,2,2,0\}$
$glb(ABC) = \max lb(ABC) = 2$

$supp(ABC) \leq s_{AB} = 4$
$\leq s_{AC} = 2$
$\leq s_{BC} = 4$
$\leq s_{AB} + s_{AC} + s_{BC} - s_A - s_B - s_C + s_\emptyset = 4 + 2 + 4 - 4 - 6 - 4 + 6 = 2$

$ub(ABC) = \{4,2,4,2\}$
$lub(ABC) = \min ub(ABC) = 2$

$glb(ABC) = lub(ABC) = 2$
and hence ABC is derivable.

# Conclusions

Association rules tell us which items we will probably see given that we've seen some other items

- many business and scientific applications

Frequent itemsets tell which items appear together

- mining these is the first step for mining many other things
  - many different algorithms exist for efficient frequent itemset mining

The number of frequent itemsets is usually too large

- exponential output space
- maximal, closed, and non-derivable itemsets provide a summarisation of a collection of frequent itemsets

# *Thank you!*

Association rules tell us which items we will probably see given that we've seen some other items

- many business and scientific applications

Frequent itemsets tell which items appear together

- mining these is the first step for mining many other things
  - many different algorithms exist for efficient frequent itemset mining

The number of frequent itemsets is usually too large

- exponential output space
- maximal, closed, and non-derivable itemsets provide a summarisation of a collection of frequent itemsets