University of Alberta

Database Laboratory

Relevance of Counting in Data Mining Tasks

Osmar R. Zaïane 蔡頤安

http://www.cs.ualberta.ca/~zaiane/

zaiane@cs.ualberta.ca

The First International Conference on Advanced Data Mining and Applications

Wuhan, China, July 22-24, 2005

Calculateur/Ordinateur

计算机

Rechner

вычислительная машина

الآلة الحاسبة

Computer

Calcolatore

الحـــسوب

---

123 4567 89012 34 567890123 456789 0 123 45678901 23 4567 890
1234 56789 012345 67 890 1234 56 78901 234 567890 12 34567 8901
2345 678 901 234 5678 90 12 3456 789 0123 456 78901 2345 678 901
23456 789 012 345 678 90123 4567 89012 34567890 12 345678 9012
34 567 89012 345 678 901 23456 7890 1 23 45 67890 123 456 7890
12345 678 90 12 345 678 901 2345 67890 123 4567 890 1234 56 78901
23 4567 890 12 3456 78 9012 3456 78 9012 345 678 90 1234 567 8901
234 567 8901 23 456 78 901 234 5678 901 234 5678 90 12 345 678
9012 3456 7890 123 4567 890 12 3456 78 901 2345 678 90 1234 56
7890 123 456 7890 12 34567 890 123 456 7890 1234 5678 9012 34 56
78 90 12 345 678 901 234 5678 9012 34 56 7890 1234 567 890 1234
5678 901 234 567 890 1234 5678 9012 34 56 78 90 12 34 567 890 123
456 78 90 12 34 56 7890 1234 5678 90 12 34 56 789 12 34 567 890 123
45 67 89 12 3456 7890 12 345 678 9012 34 567 8901 23 456 7890 12
345 6789 12345 67 890 1234 56789 12345 67890 1234 5678 901 234
56 78 901 234 567 8901 2345 67890 12345 678 901 234 5678 9012 345
678 901 23 45 678 9012 34567 8901 23 45 678 90123 45 678 9012 345
678 901 2345 67 890 12345 67 890 12 345 678 901 234 56789 1 23 56

Calculateur/Ordinateur

计算机

Rechner

вычислительная машина

الآلة الحاسبة

Computer

Calcolatore

الحـــسوب

0101000100101110101001110100100111010100111001110101011101
0100111000111001110101001110011100100101010110010101000110
0100111010101001110010100011001010011101010100101010001101
0010010011100111000010010001000101010001110100101010001110
0011101010110111010010101001000111101111110011111100111011
0011100001100010011100110101100010010101010100011001010100
1000100100100001001110101001010101011101100010101001011101
0100011100100101010100100101010100101001010001001001101100
1001001101011011101001010100100011110111111001111110011101
1001110000110001001110011010110001001010110101010010101001
0100101010101010100101010011010101001001110011100011101001
0011100010101101110100101010010001111011111100111111001110
1100111000011000100111001101011000100111110111001110001011
1000011101011011010010101001000111101111110011111100111101
1001110000110001001110011010110001001100110101011000111000
1110010110101011011101001010100100011110111111001111110011
1011001110000110001001110011010110001001100010010010101010
1011010110101101110100101010010001111011111100111111001110

Calculateur/Ordinateur

计算机

Rechner

вычислительная машина

الآلة الحاسبة

Computer

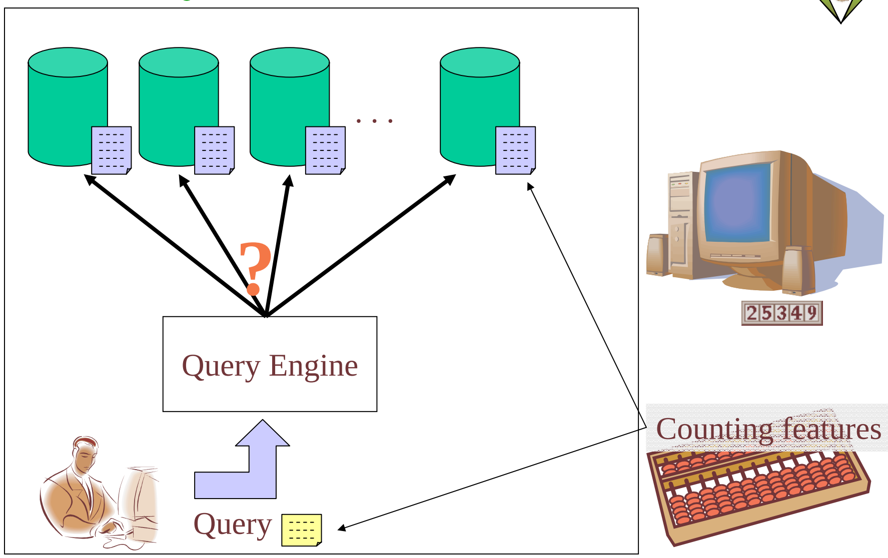
Calcolatore

الحـــسوب

University of Alberta

Database Laboratory

Database Selection Problem



University of Alberta

Database Laboratory

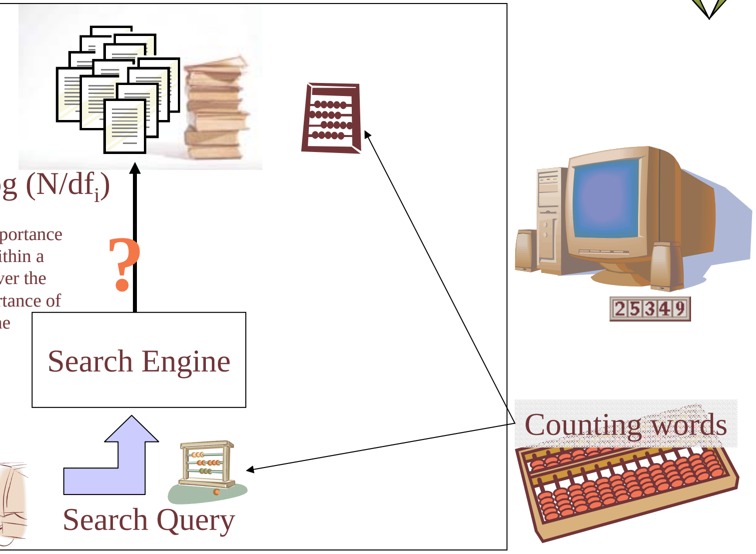
Information Retrieval

TF-idf

$W_{i,j} =$

$tf_{i,j} \times \log (N/df_i)$

Measures importance of a word i within a document j over the general importance of the word in the collection.



University of Alberta

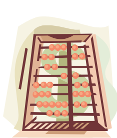
Database Laboratory

Discretization & Concept Hierarchies

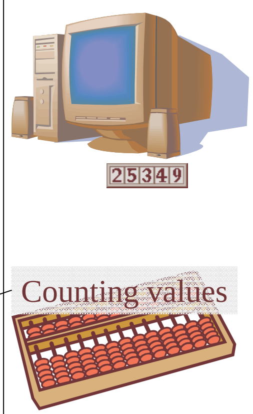
- Handling continuous data
- Automatic Concept hierarchy building
- Numerosity Reduction
- Smoothing Noise

Discretization is used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals.

Binning
Histogram analysis
Entropy-based
3-4-5 data segmentation



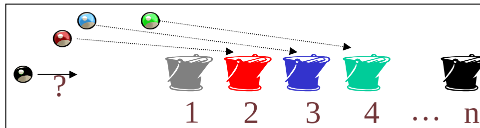
Statistics



University of Alberta

Database Laboratory

Naïve/Full Bayesian Classifier

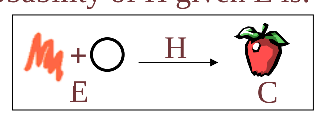


Classification = prediction for discrete and nominal values

Bayesian Learning (Bayes Theorem)

Given a hypothesis H that some data belongs to a class C and some evidence E about the data, the posteriori probability of H given E is:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

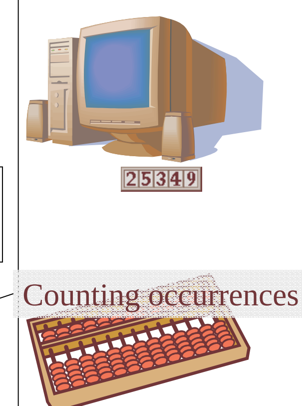


P(H)=P( )

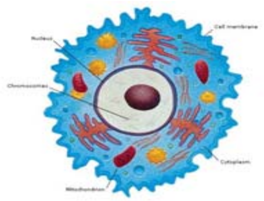
P(E) = P(+)

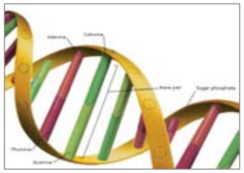
P(E|H)=P(+ if)

Learning = Calculating prior and posterior probabilities.

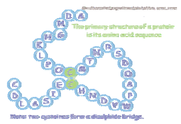


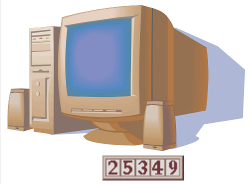
Genomics and Proteomics





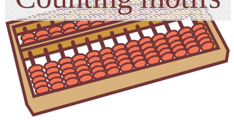
Predicting function or location of a Protein



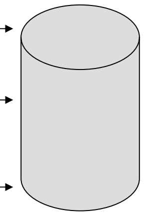


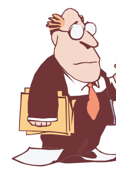
Investigate frequent sequences and sub-sequences

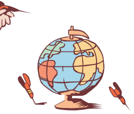
Counting motifs



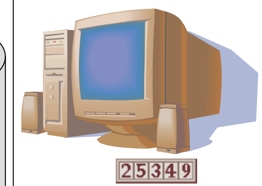
Market Basket Analysis



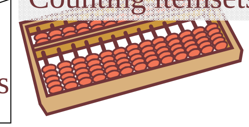




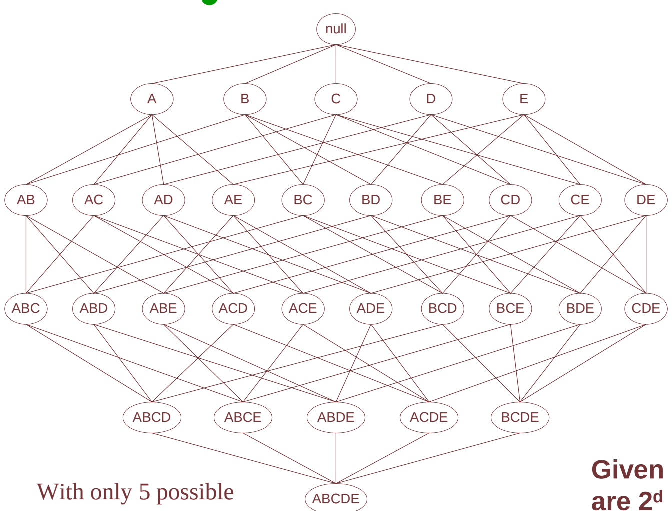
Bread, ➔ Milk
Coke, Chips ➔ Hot dogs



Counting itemsets



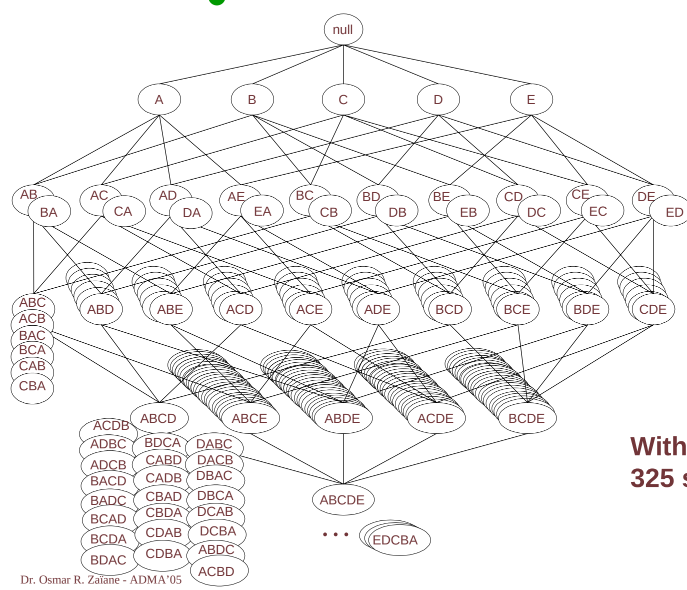
Frequent Itemset Search Space



With only 5 possible items A, B, C, D, and E, there are 2^5=32 cases.

Given d items, there are 2^d possible candidate itemsets

Frequent subsequence Search Space



With 5 items we have 325 sub-sequences

University of Alberta

Database Laboratory

Association Rule Mining

Frequent Itemset Mining

FIM 1 abc

Bound by a **support** threshold

Association Rules Generation

2 ab→c b→ac

Bound by a **confidence** threshold

- Frequent itemset generation is still computationally expensive

University of Alberta

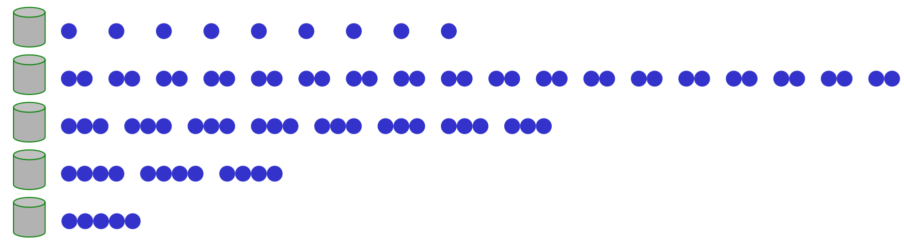
Database Laboratory

Apriori Algorithm

Apriori (Agrawal et al. 1994)

Repetitive I/O scans

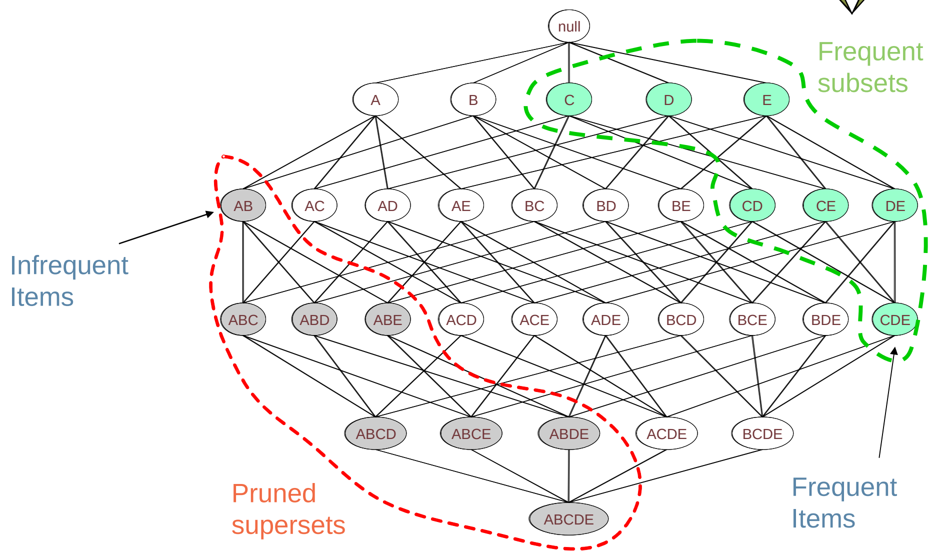
Huge Computation to generate candidate items



University of Alberta

Database Laboratory

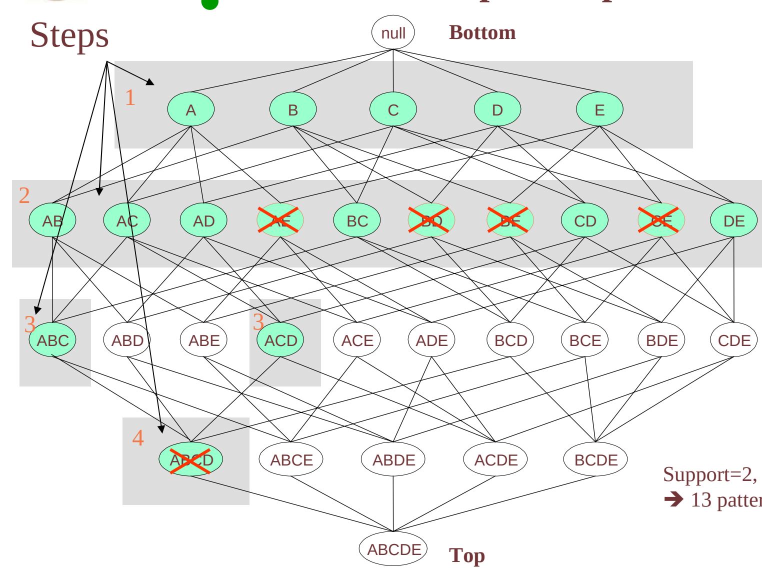
Illustrating Apriori Principle



University of Alberta

Database Laboratory

Bottom – Up Example



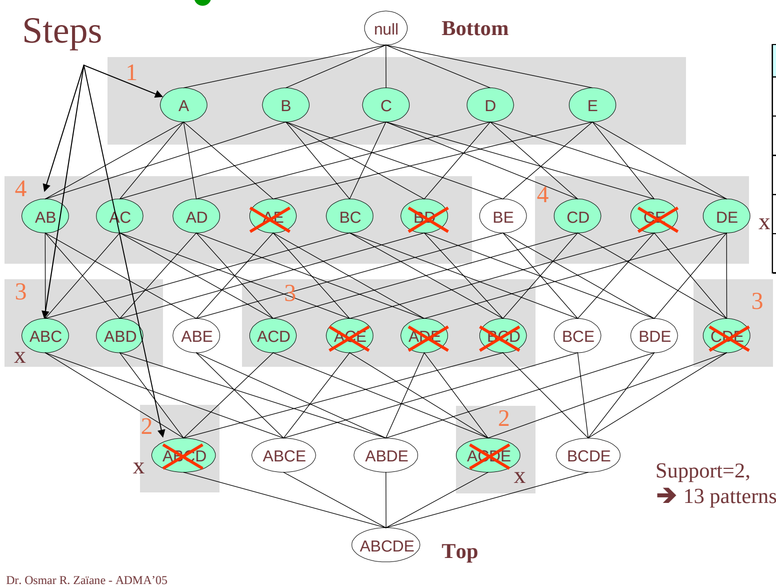
TID	Items
1	ABC
2	ABCD
3	ABC
4	ACDE
5	DE

Superset is candidate if ALL its subsets are frequent

Support=2,
➔ 13 patterns

18 candidates to check

Top - Down Example



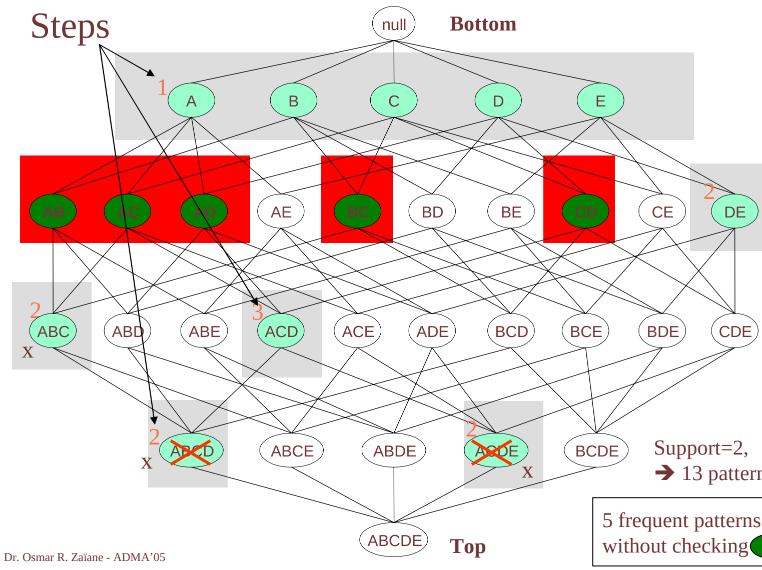
TID	Items
1	ABC
2	ABCD
3	ABC
4	ACDE
5	DE

Subset is candidate if it is marked or if one of its supersets is candidate

Support=2,
→ 13 patterns

23 candidates to check

Leap Traversal Example



TID	Items
1	ABC
2	ABCD
3	ABC
4	ACDE
5	DE

Itemset is candidate if it is marked or if it is a subset of more than one infrequent marked superset

Support=2,
→ 13 patterns

5 frequent patterns without checking

10 candidates to check

Many Candidates – Many Patterns

Not only there are too many candidate itemsets but there are also too many frequent ones.

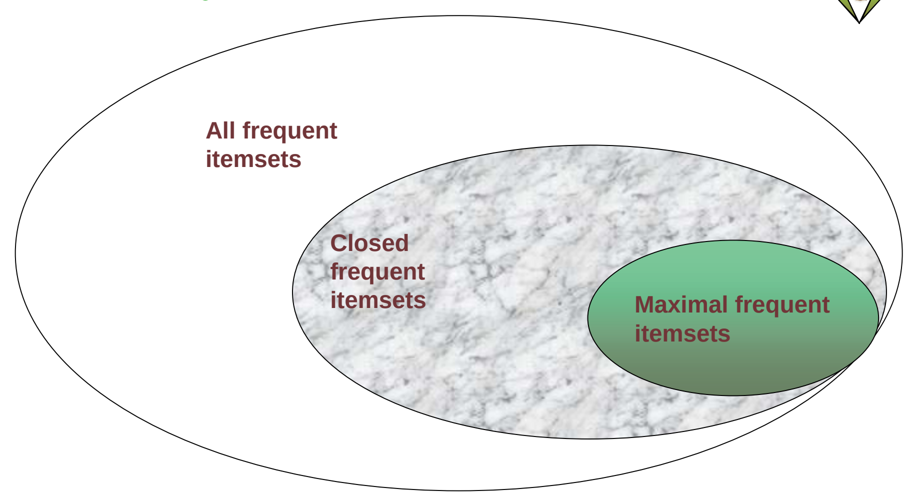
Frequent pattern $\{a_1, \ldots, a_{100}\}$

→ $(100^1) + (100^2) + \ldots + (100^{100})$

$= 2^{100}-1$

$= 1.27*10^{30}$ frequent sub-patterns!

Compressed Representation



Maximal frequent itemsets $\subseteq$ Closed frequent itemsets $\subseteq$ All frequent itemset

Frequent Closed Patterns

- N. Pasquier et al. In ICDT'99
- For frequent itemset X, if there exists no item y such that every transaction containing X also contains y, then X is a frequent closed pattern
- In other words, frequent itemset X is closed if there is no item y, not already in X, that always accompanies X in all transactions where X occurs.
- Concise representation of frequent patterns. Can generate all frequent patterns with their support from frequent closed ones.
- Reduce number of patterns and rules

{abcd}
{abc}
{bd}
Transactions
Support = 2

a	2
b	3
c	2
d	2
ab	2
ac	2
bc	2
bd	2
abc	2

Frequent itemsets

b	3
bd	2
abc	2

Frequent Closed itemsets

Frequent Maximal Patterns

- R. Bayardo. In SIGMOD'98
- Frequent itemset X is maximal if there is no other frequent itemset Y that is superset of X.
- In other words, there is no other frequent pattern that would include a maximal pattern.
- More concise representation of frequent patterns but the information about supports is lost.
- Can generate all frequent patterns from frequent maximal ones but without their respective support.

{abcd}
{abc}
{bd}
Transactions
Support = 2

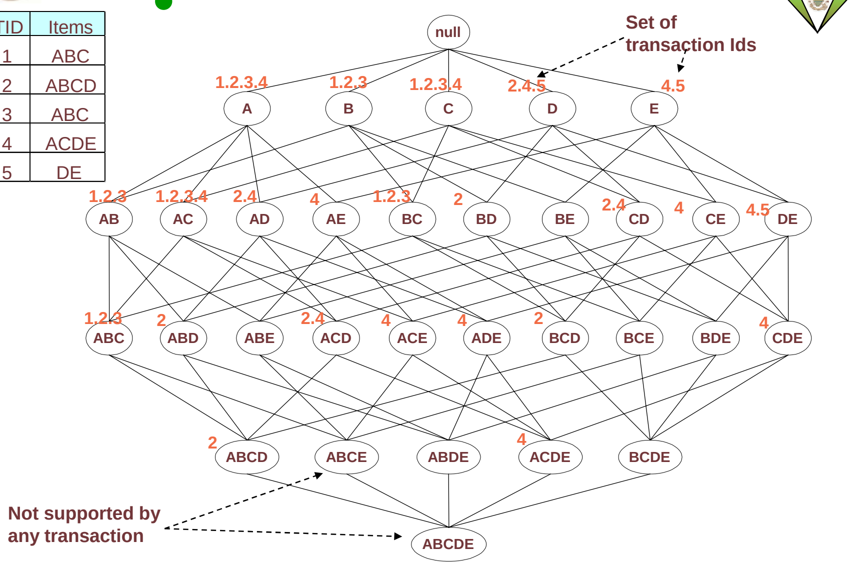
a	2
b	3
c	2
d	2
ab	2
ac	2
bc	2
bd	2
abc	2

Frequent itemsets

bd	2
abc	2

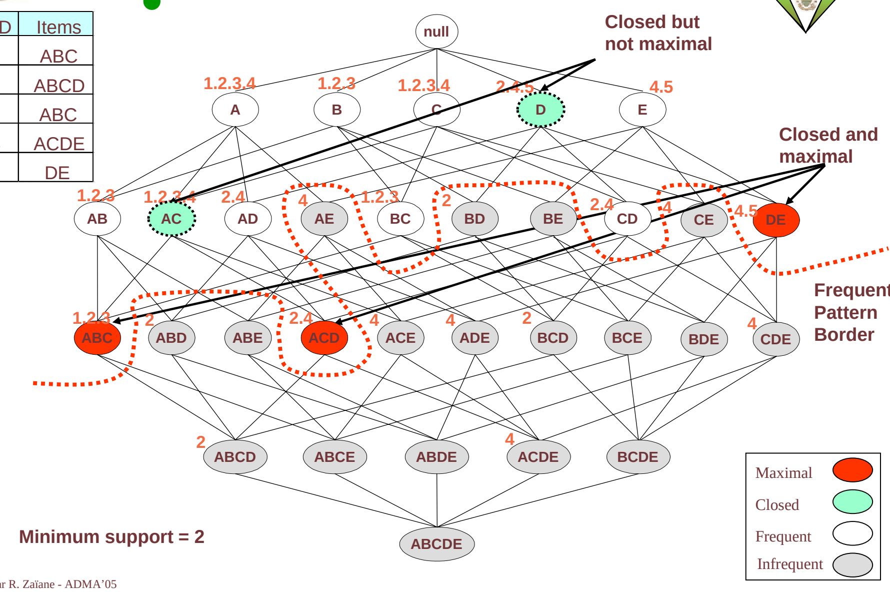
Maximal Versus Closed Patterns

TID	Items
1	ABC
2	ABCD
3	ABC
4	ACDE
5	DE

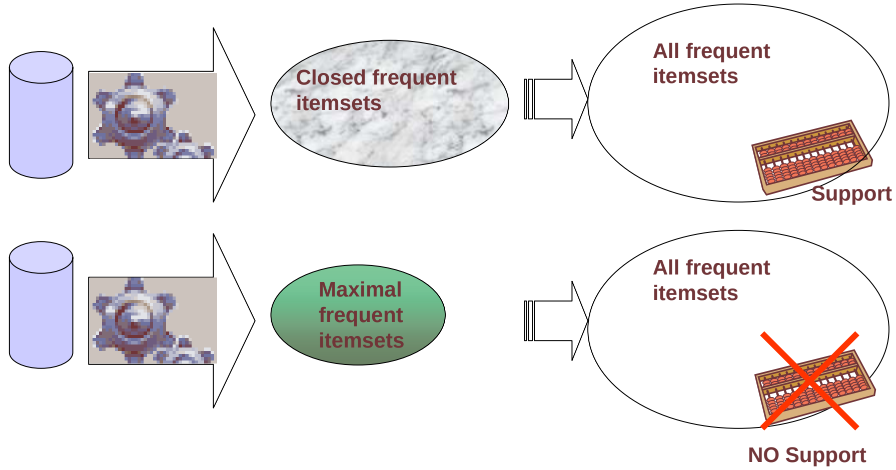


Maximal Versus Closed Patterns

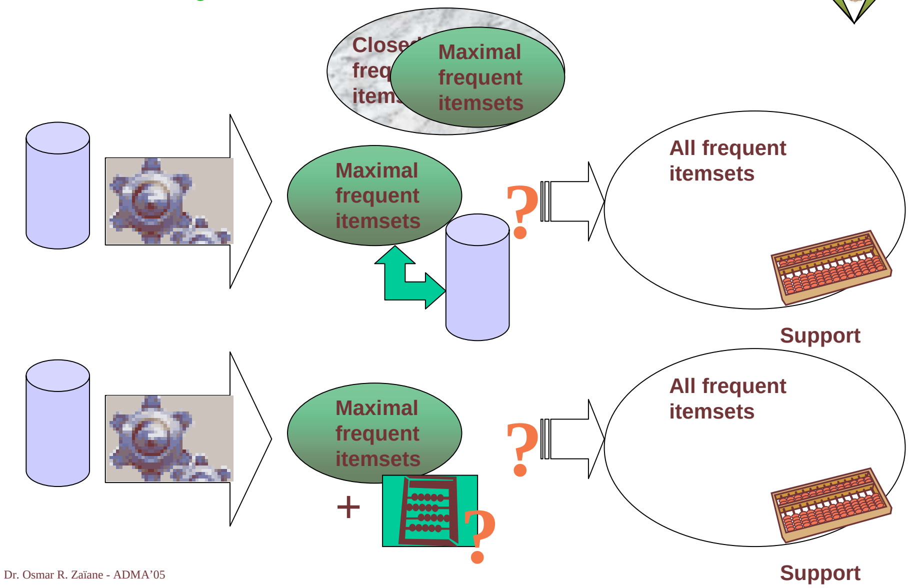
TID	Items
1	ABC
2	ABCD
3	ABC
4	ACDE
5	DE



Do We Need to Count All?



What about Maximals?

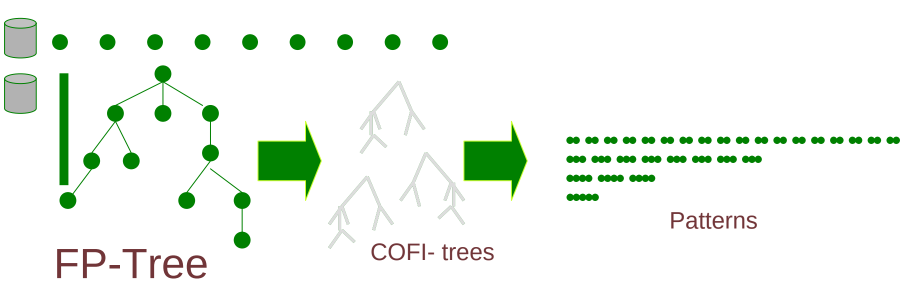


The COFI Approach

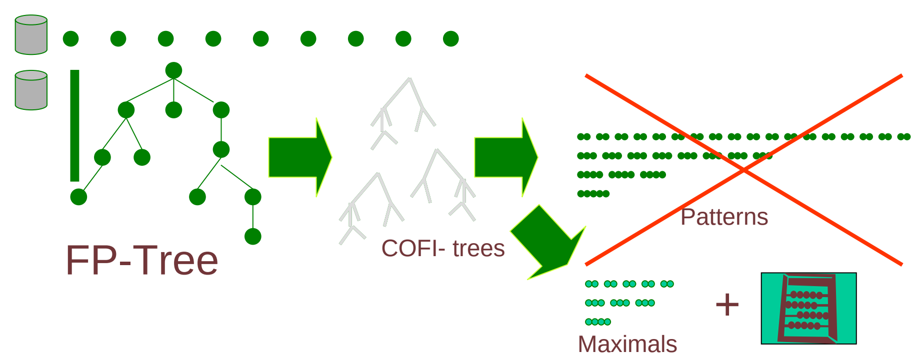
COFI

(El-Hajj and Zaïane, 2003)

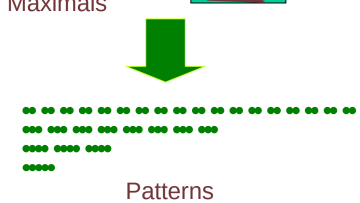
- 2 I/O scans
- reduced candidacy generation
- Small memory footprint



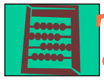
COFI-MAX and the extra info



(Zaïane and El-Hajj ACM SIGKDD 2005)



Ordered Partitioning Bases

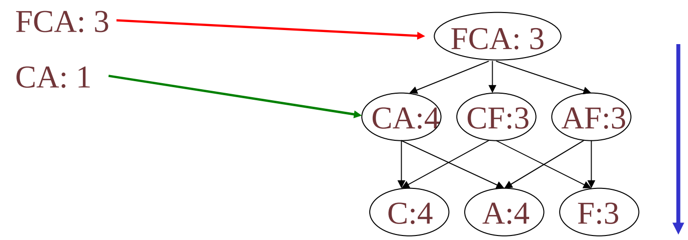
+  ?

What is this extra information?

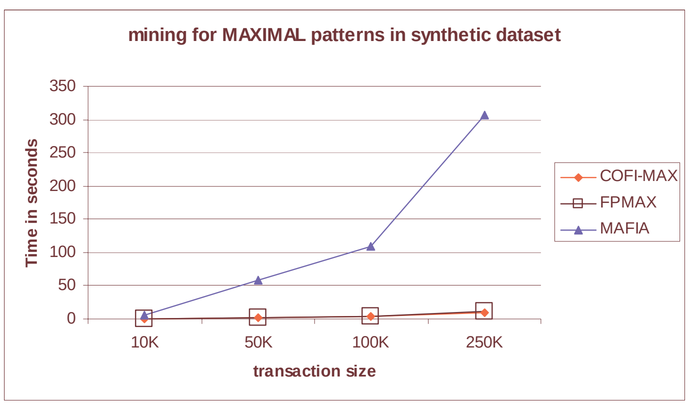
- A data structure containing *frequent pattern bases* and their *branch support*;
- The data structure is a "free" bonus since it is used to mine for maximals;
- *Frequent pattern bases* are those marked sub-transactions in the leap-approach and their descendents if not frequent.
- The *branch support* is the number of times the *frequent pattern base* occurs alone (not subsumed by another pattern)

Counting the Support

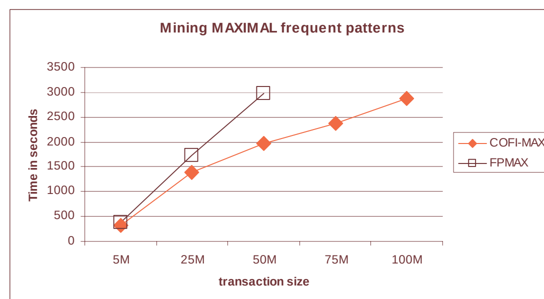
Support of any pattern is the summation of the supports of its supersets of frequent-path-bases

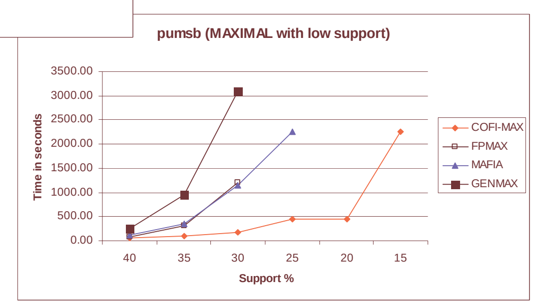


Some Results (synthetic data)



Some Other Selected Results





University of Alberta

Database Laboratory

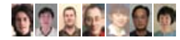
In Conclusion …

- Computers are machines that count and compute
- Many data mining tasks consist in counting
- The task of enumerating and counting is essential but not necessarily easy.
- We do not need to count all possibilities or even all patterns of direct interest
- The challenge is to reduce the enumeration without loosing effectiveness (loss-less compression)
- There is no winner / no best way to count

University of Alberta

Database Laboratory

My Students



- Maria-Luiza Antonie
- Jiyang Chen
- Mohammad El-Hajj
- Andrew Foss
- Yan Jin
- Yi Li
- Yaling Pei

- Stanley Oliveira
- Yang Wang
- Lisheng Sun
- Jia Li
- Alex Strilets
- William Cheung
- Yue Zhang
- Chi-Hoon Lee
- Weinan Wang
- Ayman Ammoura
- Hang Cui
- Jun Luo
- Yuan Ji

Without my students this research work wouldn't have been possible.

University of Alberta

Database Laboratory

谢谢

蔡頤安

UNIVERSITY OF ALBERTA

Osmar R. Zaïane, Ph.D.
Associate Professor
Department of Computing Science

221 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: zaiane@cs.ualberta.ca
http://www.cs.ualberta.ca/~zaiane/

http://www.cs.ualberta.ca/~zaiane/