Associação (Cesta de Compras)

Association Rules
Regras de Associação

Regras de Associação

- Um grande conjunto de itens, por exemplo, as coisas vendidas em um supermercado.
- Um grande conjunto de cestas, cada um dos quais é um pequeno conjunto de itens, por exemplo, a coisas que um cliente compra em uma visita ao supermercado.
- Um mapeamento geral (associação de muitos-para-muitos) entre dois tipos de coisas, onde uma coisa (cestas) é um conjunto de outras coisas (os itens).
- Porém o foco são as conexões entre os "itens", não entre as "cestas".
- A técnica se concentra em eventos comuns, e não em eventos raros.

O Problema

- O problema dos conjuntos frequentes de itens é o de encontrar conjuntos de itens que aparecem em muitas cestas.

Análise de Associação

- Uma aplicação típica é a análise do comportamento de compra do consumidor em lojas (supermercados ou redes), onde registram o conteúdos dos carrinhos de compras (cestas) levados ao caixa.
- Esses dados de transação são normalmente registrados por scanners de ponto de venda e consistem em registros no formato: {transação ID, item ID, item ID, ...}.
- Ao encontrar conjuntos de itens frequentes um varejista pode aprender o que é comumente comprado em conjunto pelos consumidores e usar essas informações para aumentar as vendas de várias maneiras.

Conjuntos de Itens Frequentes

- Dado um conjunto de transações, descubra a combinação de itens (conjuntos de itens – ou *itemsets*) que ocorrem frequentemente.

Suporte $s(I)$: número de transações que contém o itemset I

Market-Basket transactions

Items: {Bread, Milk, Diaper, Beer, Eggs, Coke}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Exemplos de *itemsets* frequentes $s(I) \geq 3$

{Bread}: 4
{Milk} : 4
{Diaper} : 4
{Beer}: 3
{Diaper, Beer} : 3
{Milk, Bread} : 3

Suporte e Confiança

$$\text{SUPPORT} = \frac{\text{number of transactions containing X and Y}}{\text{total number of transactions}}$$

$$\text{CONFIDENCE} = \frac{\text{number of transactions containing X and Y}}{\text{number of transactions containing X}}$$

- Regra de Associação
  - Uma expressão de implicação na forma $X \rightarrow Y$, onde X e Y são conjuntos de itens.
  - Exemplo:
    {Milk, Diaper} → {Beer}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Métricas para Avaliação
  - Suporte (s)
    - Fração de transações que contêm X e Y. Nos diz o quão popular é o conjunto de itens.
  - Confiança (c)
    - Mede a frequência com que os itens em Y aparecem em transações que contêm X. Probabilidade de Y em transações que contenham X.

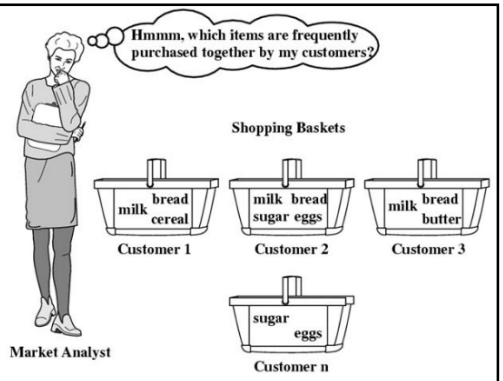
Exemplo:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Aplicações



- Cestas = conjuntos de produtos que alguém comprou em uma visita à uma loja.
- Itens = produtos que podem estar contidos nas cestas.
- Exemplo de aplicação: uma vez que muitas pessoas compram cerveja e fraldas juntas:
  - Diminuir o preço de fraldas e aumentar o preço da cerveja ou vice-versa.
  - Colocar fraldas e cervejas em gondolas próximas ou induzir o comprador de fraldas passar pela gondola da cerveja ou vice-versa.
- Só será vantajoso desde que muitos clientes comprem fraldas e cerveja.

Outras Aplicações

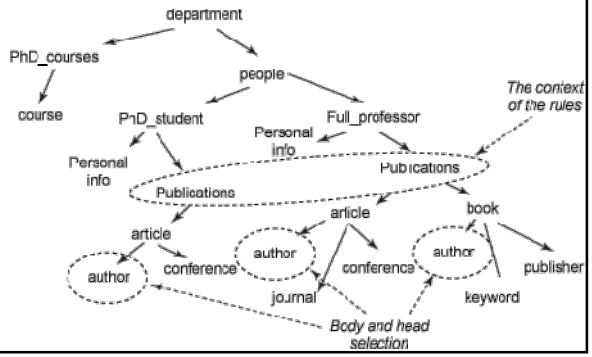
- Cestas = páginas Web.
- Itens = palavras.

- Exemplo de aplicação:
  - Palavras pouco comuns aparecem em conjunto em um grande número de documentos:
  - Por exemplo, "Brad" e "Angelina," pode indicar uma relação interessante.
  - Por exemplo, “Democracia”, “Liberdade” e “Fake”, pode indicar uma relação interessante.

Aplicações

- Cestas = frases.
- Itens = documentos que contêm essas frases.

- Exemplo de aplicação:
  - Os itens que aparecem juntos com muita frequência podem representar plágio.
- Note que:
  - Os itens não precisam estar “dentro” das cestas.



Definição: Conjunto de Itens Frequente

- Conjunto de itens (Itemset)
  - Uma coleção de um ou mais itens
    - Exemplo: {Milk, Bread, Diaper}
  - K-itemset
    - Um itemset que contém k itens

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Suporte (σ)
  - Contagem: Frequência de ocorrência de um itemset
  - Por exemplo. σ({Milk, Bread, Diaper}) = 2
  - Fração: fração das transações que contêm um itemset
  - Por exemplo. s({Milk, Bread, Diaper}) = 40%
- Conjunto de Itens (Itemset) Frequente
  - Um conjunto de itens cujo Suporte é maior ou igual a um limiar (minsup) de análise.

$$s(I) \geq \text{minsup}$$

Mineração de Itemsets Frequentes

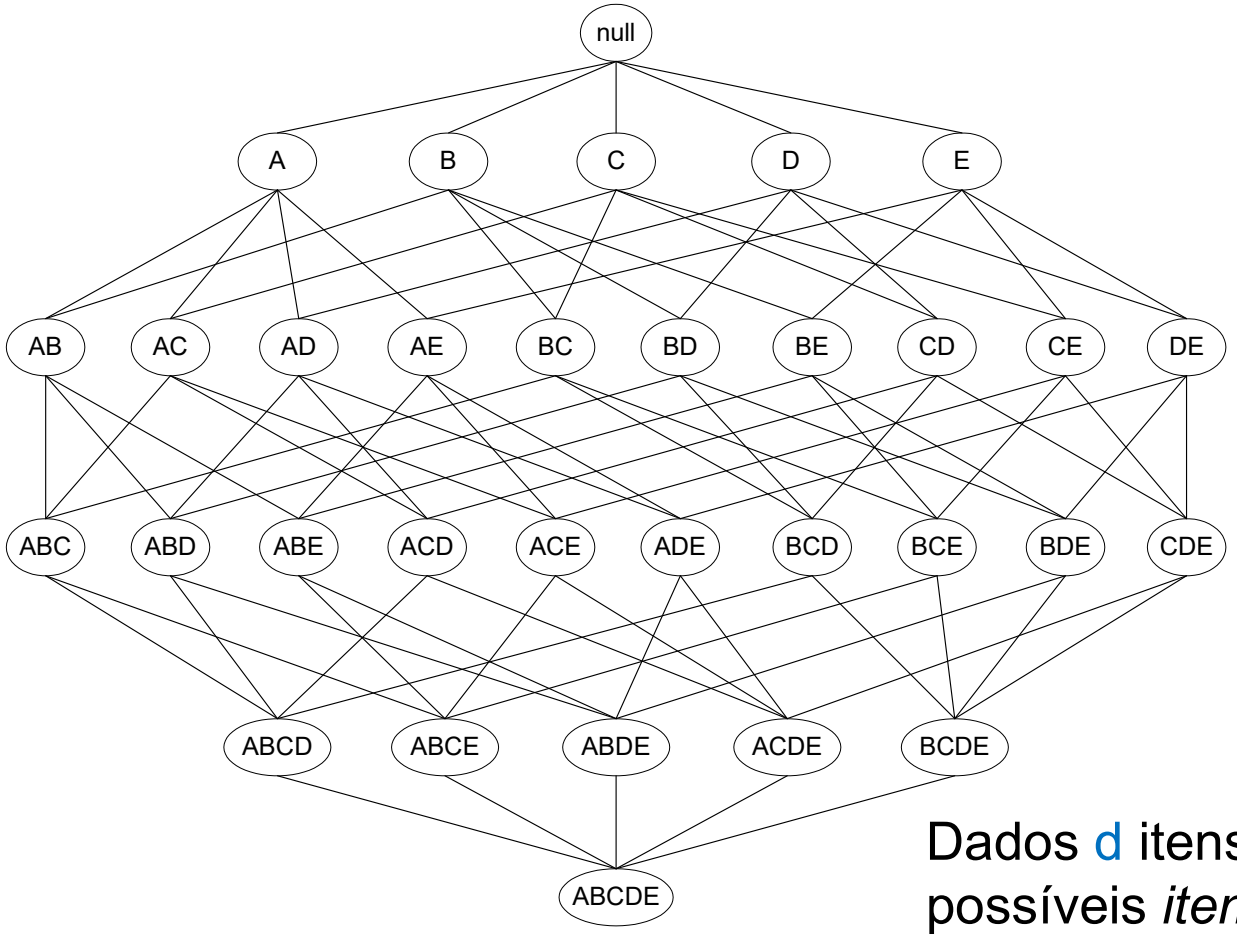
- Entrada: Um conjunto de transações T, ao longo de um conjunto de itens I
- Saída: Todos os conjuntos de itens com itens em I tendo:
  - Suporte ≥ minsup
- Parâmetros do problema:
  - N = | T |: número de transações
  - d = | I |: número de itens (distintos)
  - w: “tamanho máximo” de uma transação
  - Número de possíveis itemsets? $M = 2^d$
- Exemplos de escala do problema:
  - O WalMart comercializa 100.000 itens e pode vender milhões de cestas.
  - A Web tem muitos bilhões de palavras e muitos milhões de páginas.

A Tarefa de Descobrir Associações

- Dado um conjunto de transações T, o objetivo da análise de regras de associação é encontrar todas as regras que tenham:
  - Support ≥ *minsup*
  - Confidence ≥ *minconf*
- Abordagem de força bruta:
  - Listar todas as regras de associação possíveis
  - Calcular o suporte e a confiança para cada regra
  - Desconsiderar regras abaixo de *minsup* e *minconf*

⇒ Computacionalmente proibitivo!

A Rede de Possíveis Conjuntos de Itens



Dados d itens, há 2^d possíveis *itemsets*

Princípio Apriori

- Princípio Apriori (observação principal):
- Se um conjunto de itens é frequente, então, todos os seus subconjuntos também devem ser frequentes.
- Se um conjunto de itens não é frequente, então, todos os seus super conjuntos não podem ser frequentes.

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

- O suporte de um conjunto de itens nunca excede o suporte dos seus subconjuntos.
- Isto é conhecido como a propriedade anti-monotônica da medida de suporte.

AGV3

Principio Apriori

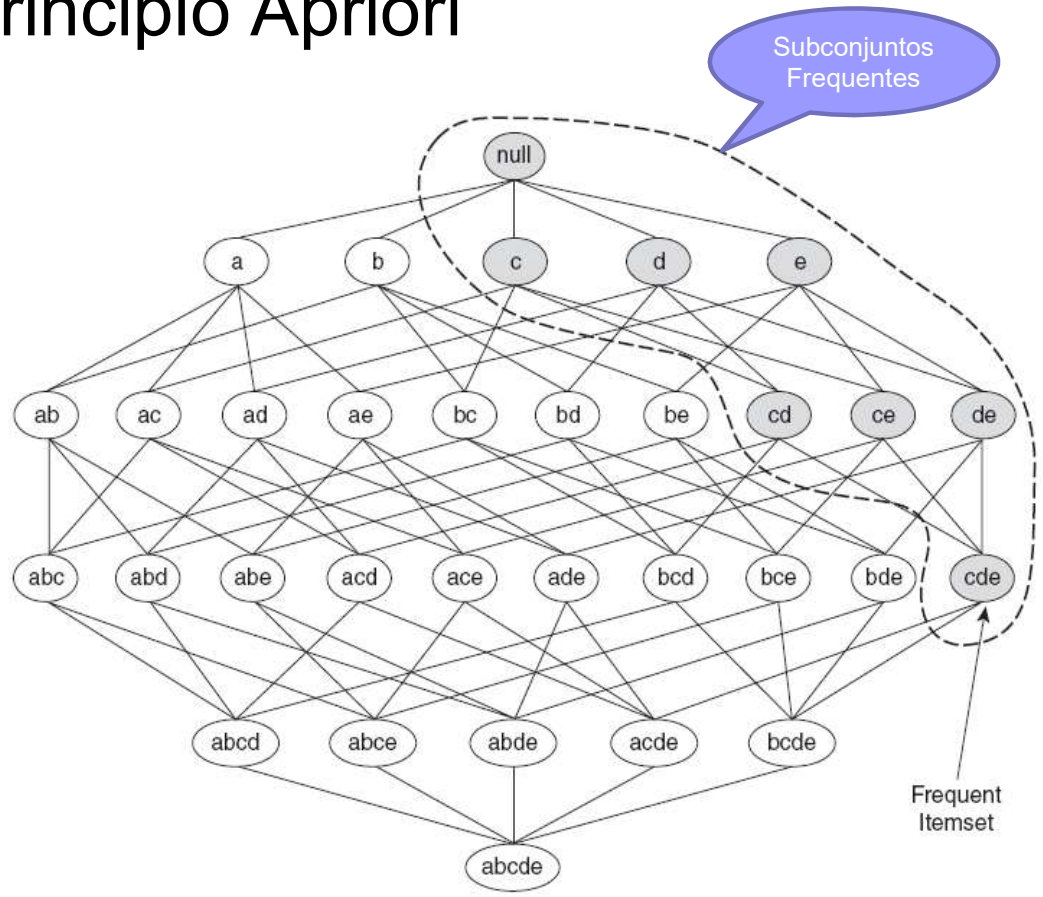
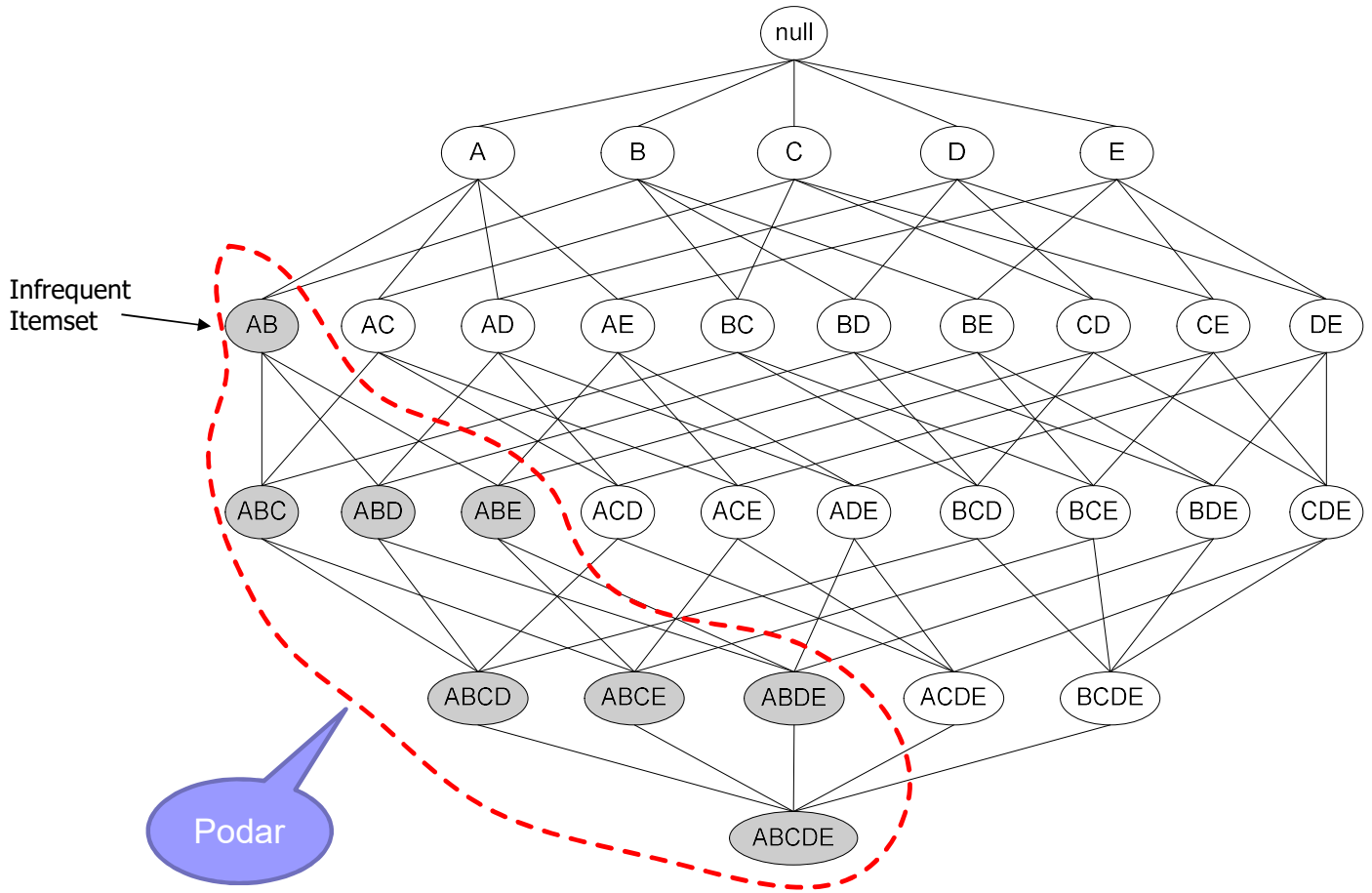


Figure 6.3. An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.

AGV3 Antonio Geraldo Vidal; 06/08/2016

Princípio Apriori



Algoritmo Apriori

C_k = itemsets candidatos de tamanho k
L_k = itemsets frequentes de tamanho k

1. k = 1, C_1 = todos os itens
2. Enquanto C_k não for vazio

Geração de Itens Frequentes

3. Percorrer o banco de dados para descobrir quais itemsets em C_k são frequentes e colocá-los em L_k

Geração de Candidatos

4. Usar L_k para gerar uma coleção de itemsets candidatos C_{k+1} de tamanho k+1
5. k = k+1

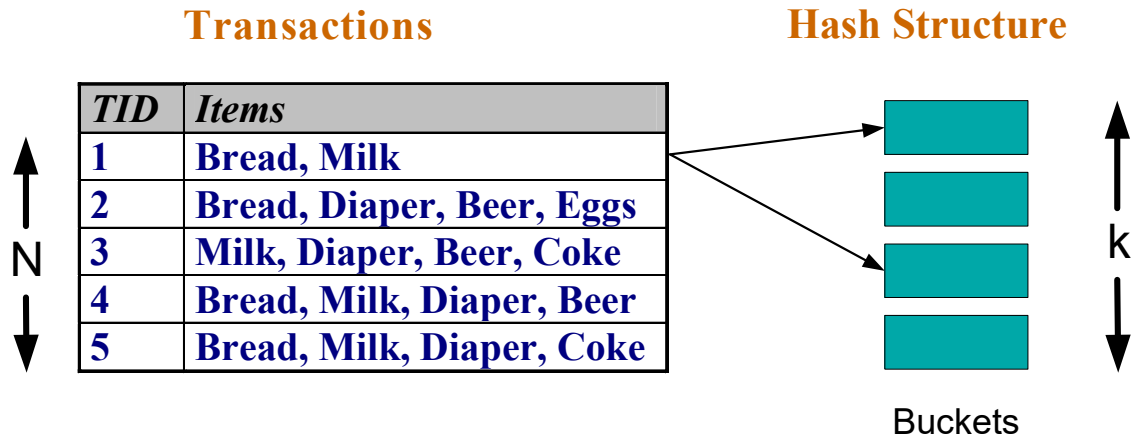
R. Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules",
Proc. of the 20th Int'l Conference on Very Large Databases, 1994.

Geração de Candidatos

- Princípio básico (Apriori):
  - Um conjunto de itens de tamanho k + 1 é candidato a ser frequente somente se todos os seus subgrupos de tamanho k forem frequentes.
- Ideia principal:
  - Construir um candidato do tamanho k + 1 pela combinação de dois conjuntos de itens frequentes de tamanho k.
  - Podar os K+1 itemsets gerados que não têm todos os k-subconjuntos frequentes.

Computando Itens Frequentes

- Dado o conjunto de itemsets candidatos C_k, precisamos calcular o suporte e encontrar os conjuntos de itens frequentes L_k.
- Percorrer os dados e usar uma estrutura de dicionário para manter um contador para cada conjunto de itens candidato que aparecer nos dados.



Um Dicionário Simples

- Criar um dicionário (*tabela hash*) que armazena os itemsets candidatos como chaves, e o número de aparições como o valor.
- Inicializar com um zero.
- Incrementar o contador para cada conjunto de itens que você encontrar nos dados.

Fatores que afetam a Complexidade

- Escolha do limiar mínimo de suporte
  - Abaixar o limiar de suporte resulta em mais itemsets frequentes
  - Isso pode aumentar o número de candidatos e comprimento máximo de conjuntos de itens frequentes
- Dimensionalidade (número de itens) do conjunto de dados
  - Mais espaço para armazenar contagem de suporte de cada item
  - Se o número de itens frequentes também aumenta, tanto os custos de computação como de armazenamento também podem aumentar
- Tamanho da base de dados
  - Desde que o Apriori faz várias passagens, o tempo de execução do algoritmo pode aumentar com o número de transações
- Largura média de transações
  - Aumenta com conjuntos de dados mais densos
  - Isso pode aumentar o comprimento máximo de conjuntos de itens e gerar *hash trees* transversais.

Geração da Regra de Associação

- Dado um conjunto de itens frequentes L, encontrar todos os subconjuntos não vazios de $f \subset L$ tal que $f \rightarrow L - f$ satisfaz o requisito mínimo de confiança
- Se {A,B,C,D} é um itemset frequente, regras candidatas:

ABC →D,	ABD →C,	ACD →B,	BCD →A,
A →BCD,	B →ACD,	C →ABD,	D →ABC
AB →CD,	AC → BD,	AD → BC,	BC →AD,
BD →AC,	CD →AB,		

- Se |L| = k, então há $2^k - 2$ regras de associação candidatas (ignorando $L \rightarrow \varnothing$ e $\varnothing \rightarrow L$).

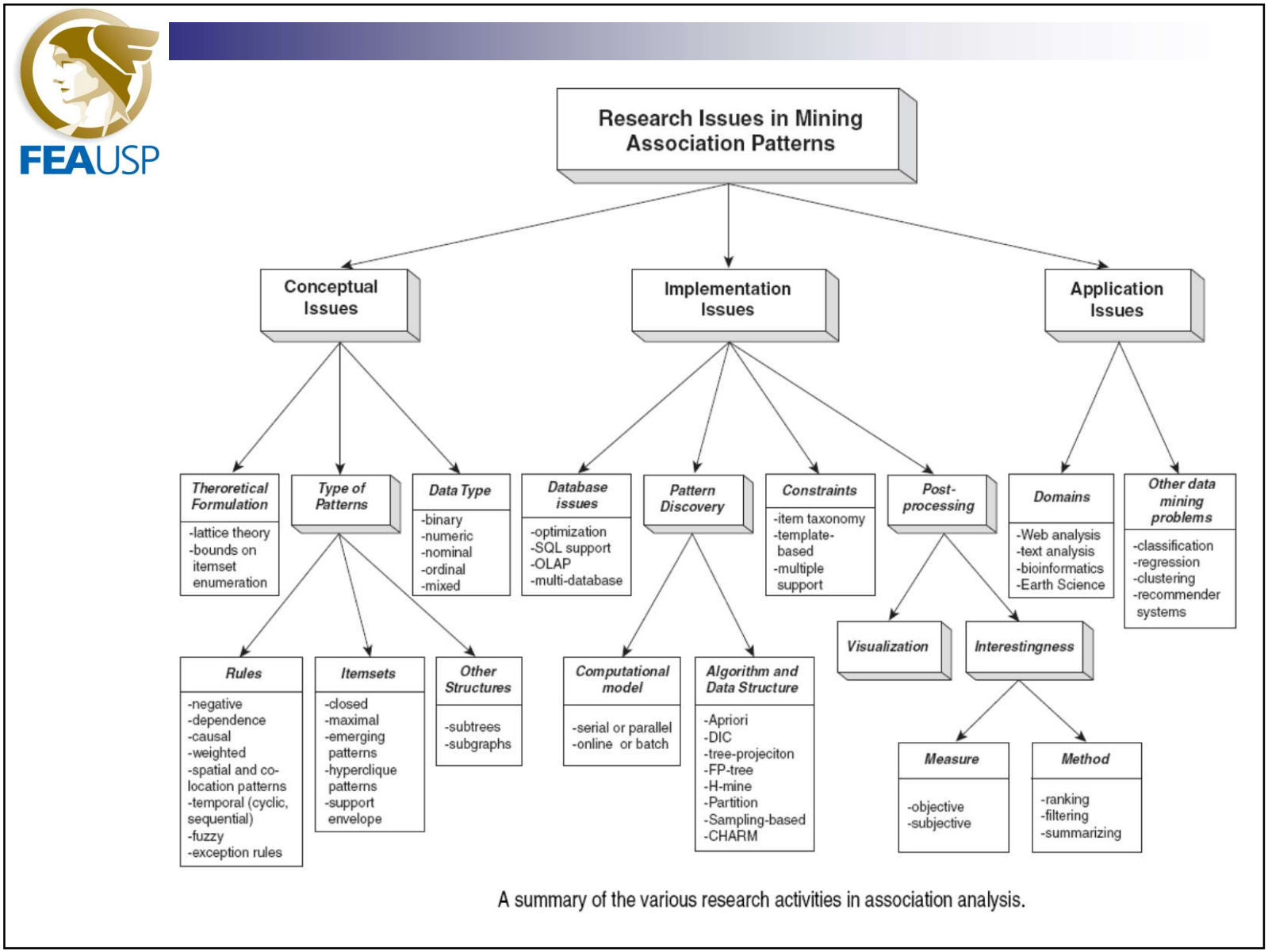
Definição do Suporte

- Como definir o limite minsup apropriado?
- Se minsup está muito alto, nós poderíamos perder itemsets envolvendo itens raros interessantes (por exemplo, produtos caros).
- Se minsup está muito baixo, é computacionalmente proibitivo e o número de conjuntos de itens será muito grande.
- Usar um único limiar mínimo de suporte pode não ser eficaz.

Avaliação do Resultado

- Algoritmos de regras de associação tendem a produzir muitas regras.
- Muitas delas são desinteressantes ou redundantes
- Redundante se {A, B, C} → {D} e {A, B} → {D}?
  Têm mesmo suporte e confiança.

- Métricas interessantes podem ser utilizadas para retirar / classificar os padrões derivados.
- Na formulação original de regras de associação, o suporte e a confiança são as únicas métricas utilizadas.



A summary of the various research activities in association analysis.