Compressão de Arquivos Orientados a Coluna com PPM

Vinicius F. Garcia[1], Sergio L. S. Mergen[1]

[1]Centro de Tecnologia
Universidade Federal de Santa Maria, Brasil

Escola Regional de Banco de Dados, Londrina, Brasil

14 de abril de 2016

Sumário

Motivação

Dados orientados a coluna

- Crescente popularização
- Abordagem NoSQL
- Consultas analíticas em poucas coluna

Compressão de dados

- Utilização econômica de memória
- Melhor aproveitamento de servidores de BD
- Necessidade de algoritmos eficientes

Dados orientados a coluna

id	*nome*	*sal.*	*inicio*
1	João Dias	2000	10-2014
2	Ana Galo	3500	05-2012
3	J. Andre	2800	12-2015

Sistema orientado a registros

$\rightarrow$

$arq_1 \Rightarrow$ 1, 2, 3
$arq_2 \Rightarrow$ João Dias, Ana Galo, J. Andre
$arq_3 \Rightarrow$ 2000, 3500, 2800
$arq_4 \Rightarrow$ 10-2014, 05-2012, 12-2015

Sistema orientado a coluna

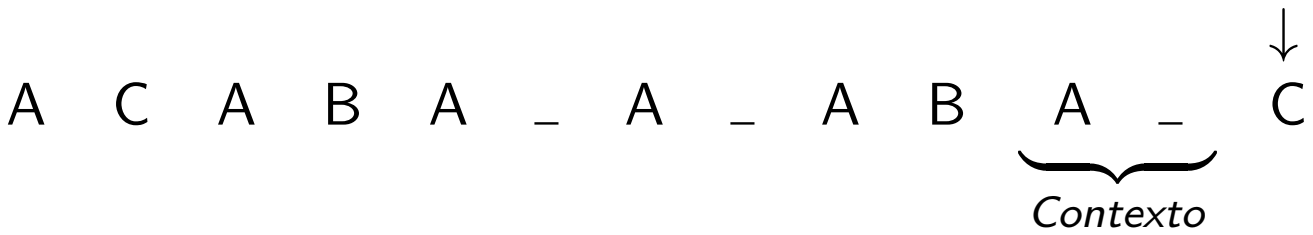
Objetivo

Determinar se algortimos da família dos PPMs (Prediction by Partial Matching) são adequados e eficientes em relação aos demais testados para dados orientados a coluna.

Metodologia do PPM

Informações gerais

- O método PPM codifica um símbolo de cada vez
- Utilização de um histórico chamado de contexto
- O primeiro símbolo após o contexto é o alvo da codificação
- A organização do contexto é tipicamente implementada com árvores
- A árvore gera um intervalo de probabilidade de aparição do símbolo



Metodologia do PPM

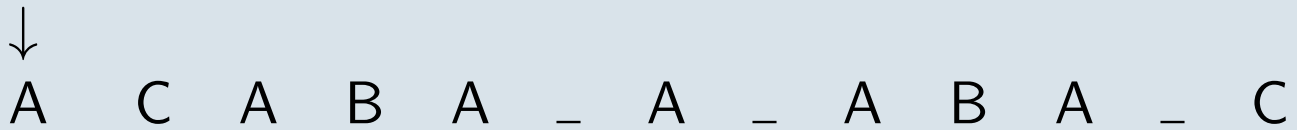
Busca e codificação

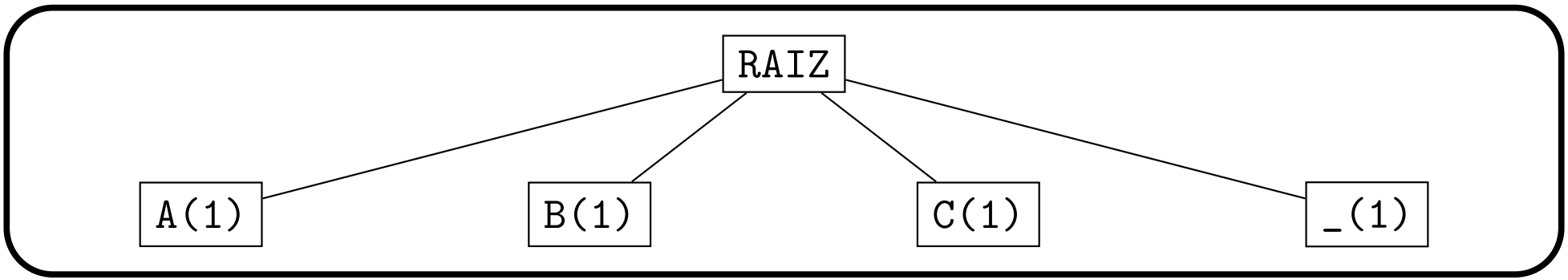
- Busca pelo símbolo alvo nível a nível que estejam disponíveis
- Níveis com símbolo alvo ausente resultam na codificação de um escape
- Entropia realizada por codificação aritimética

A C A B A _ A _ A B A _ C

Metodologia do PPM

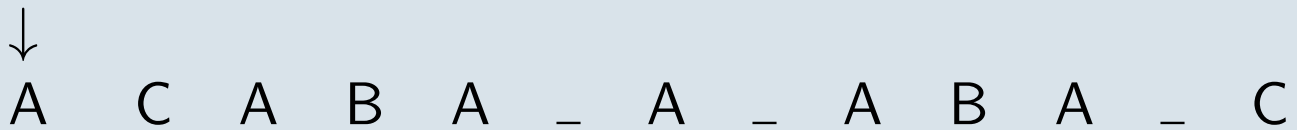
Exemplificação

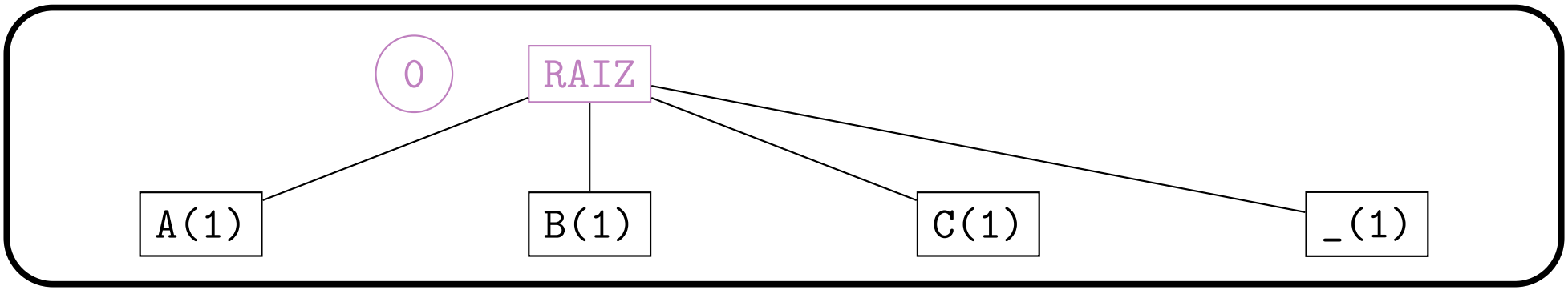




Metodologia do PPM

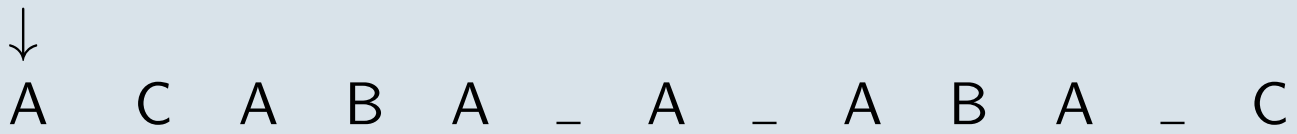
Exemplificação

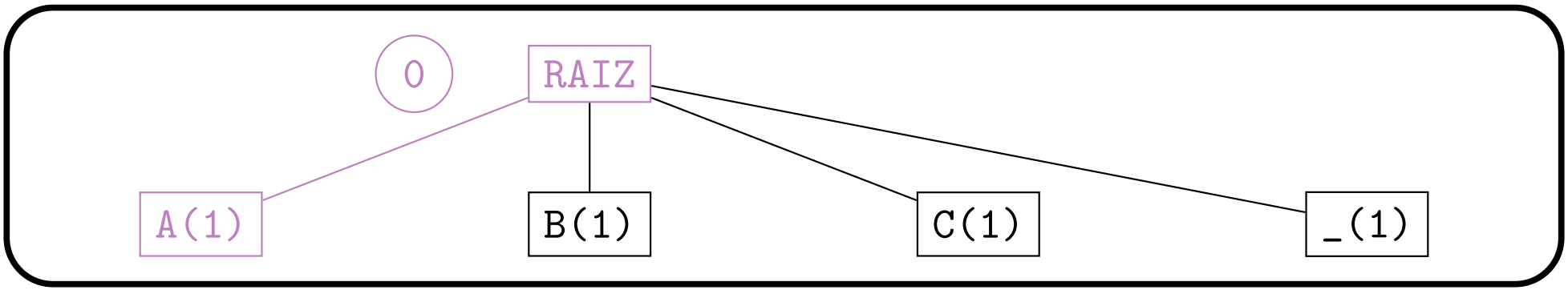




Metodologia do PPM

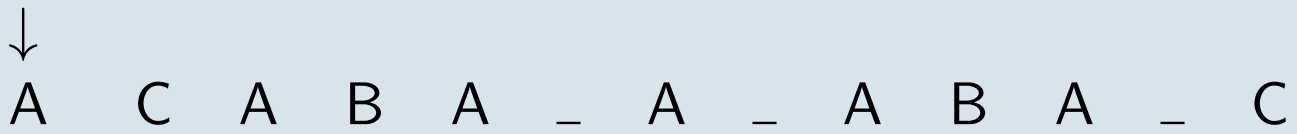
Exemplificação

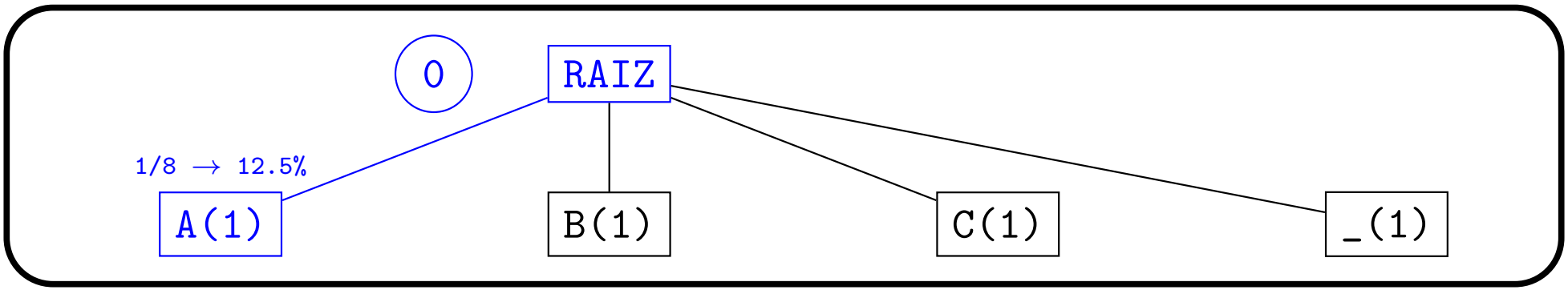




Metodologia do PPM

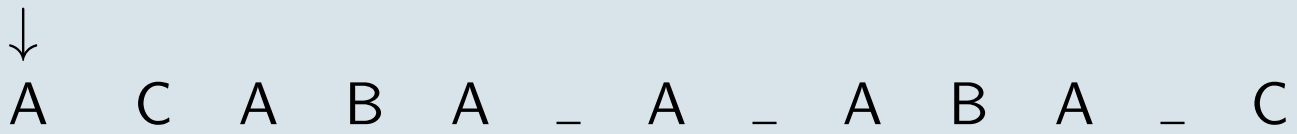
Exemplificação

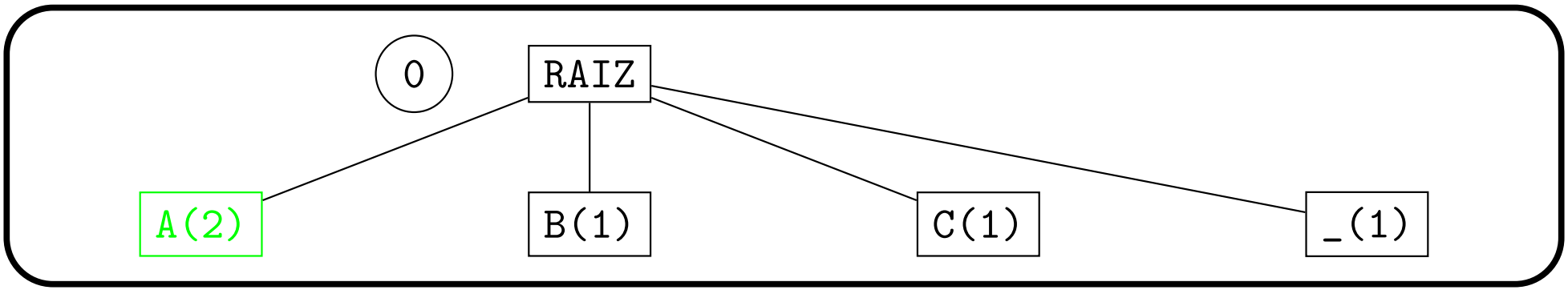




Metodologia do PPM

Exemplificação



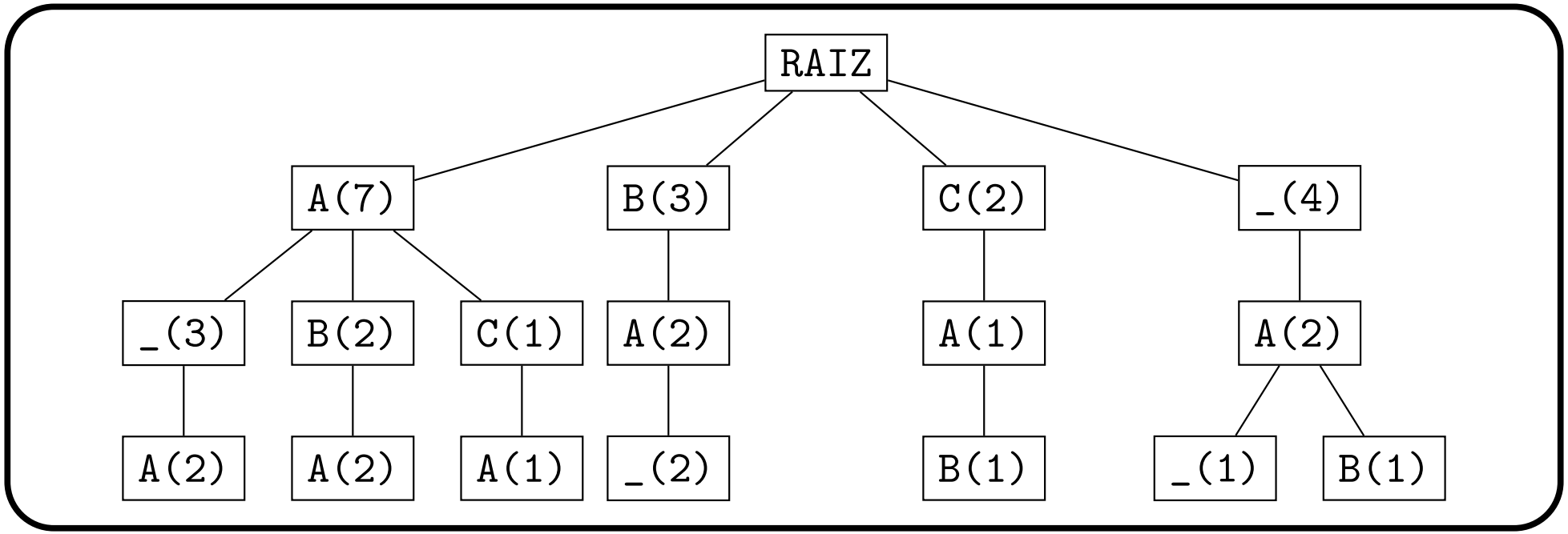


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

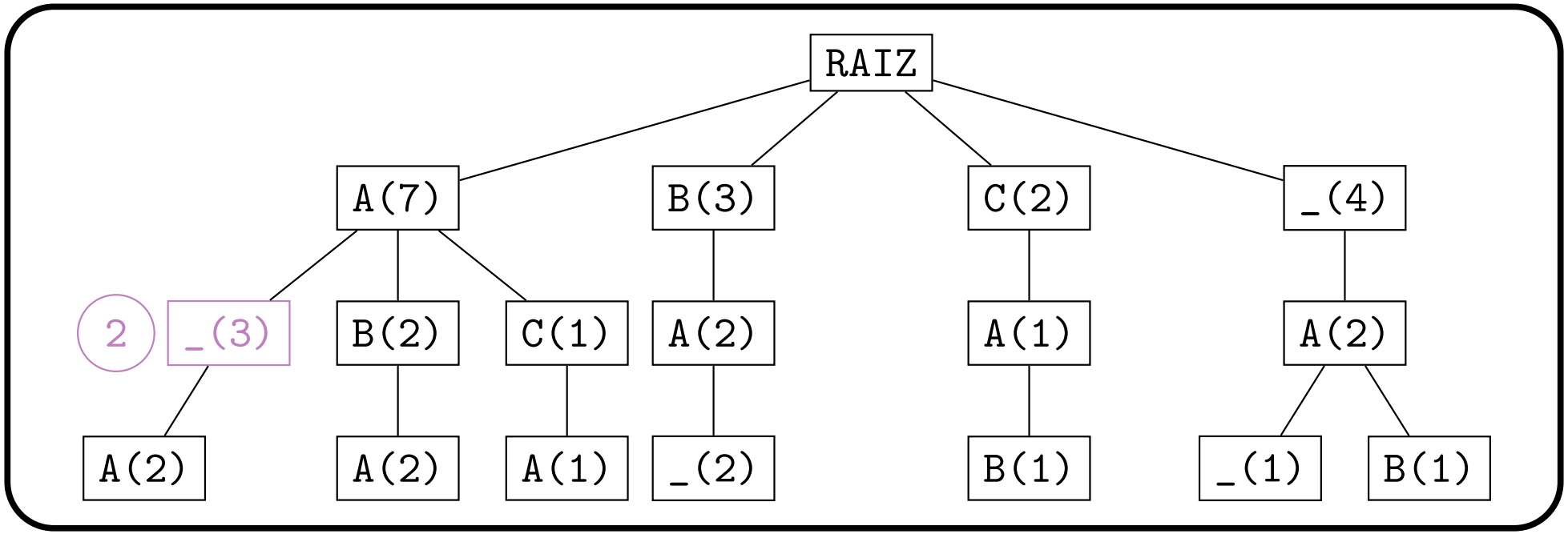


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

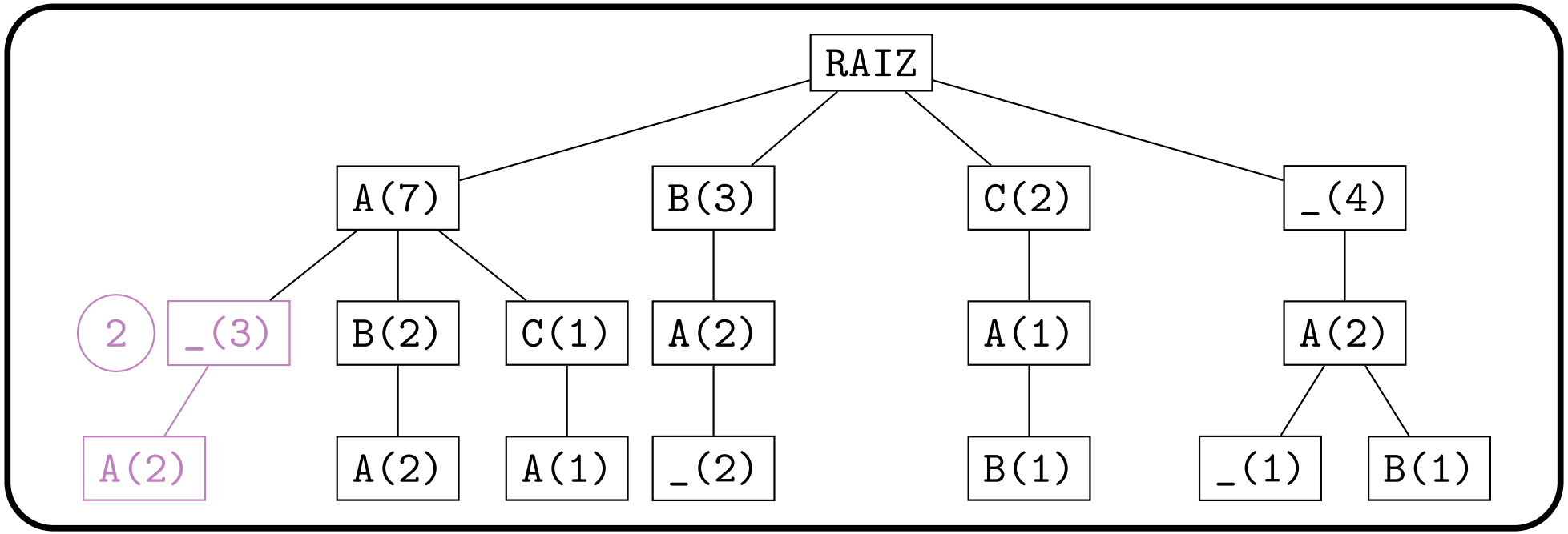


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

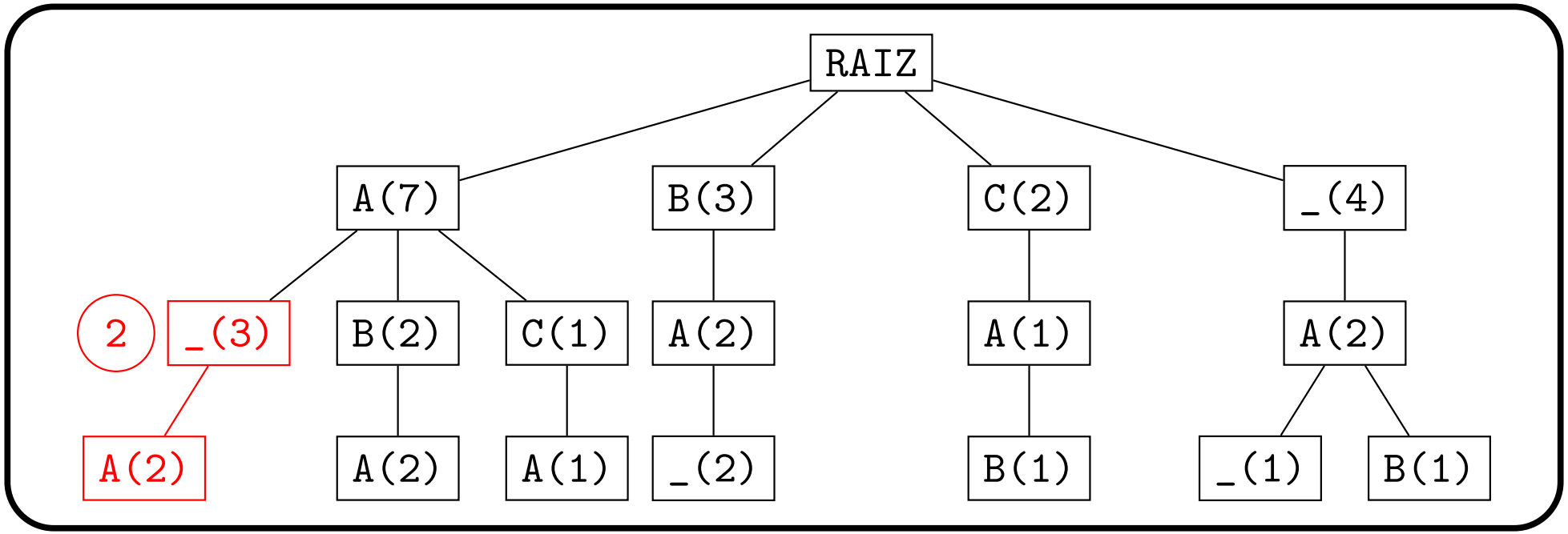


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

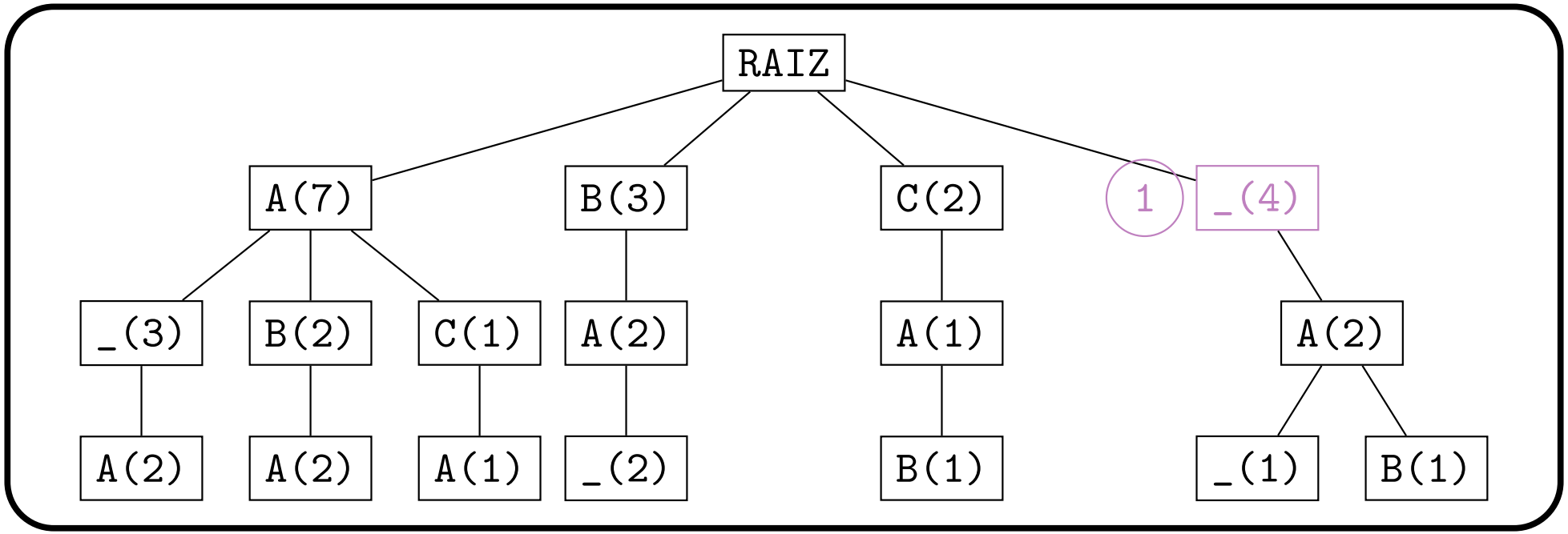


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

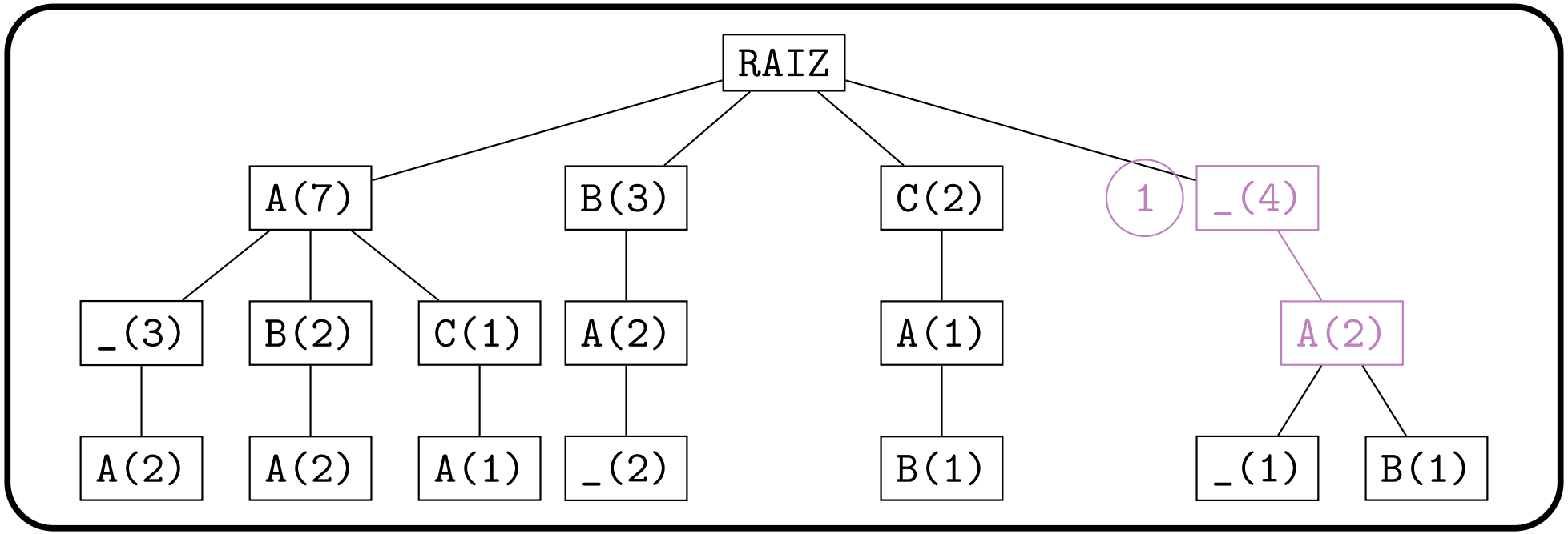


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

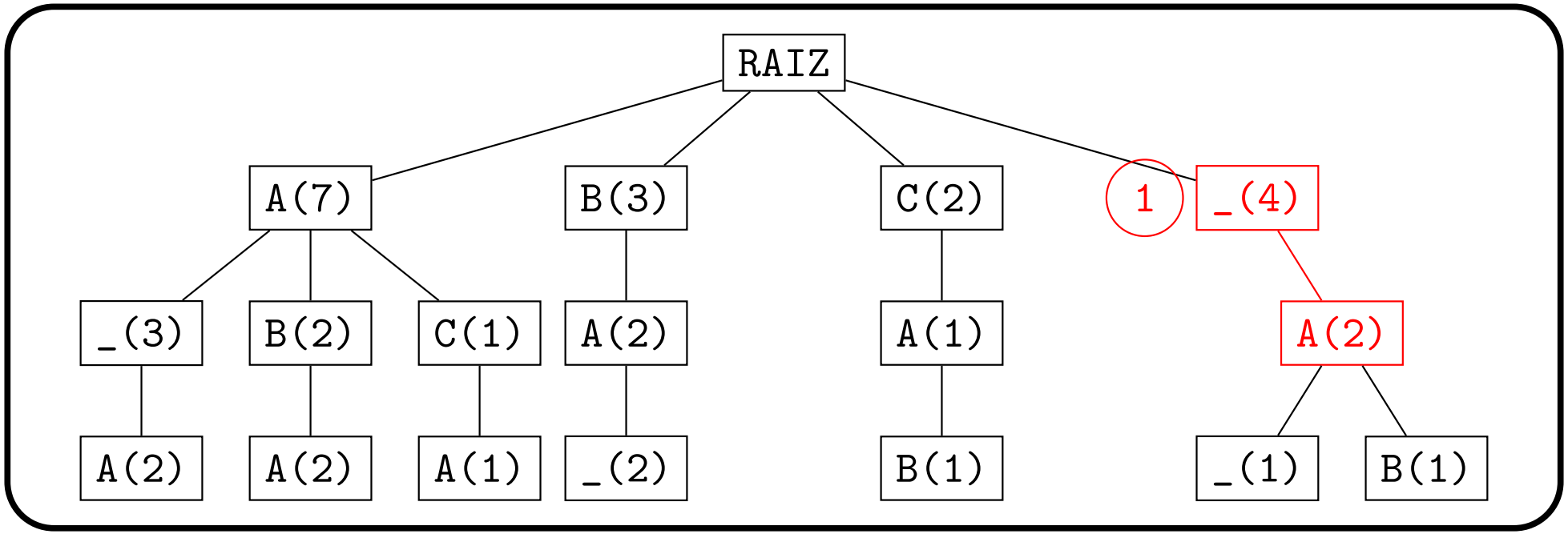


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

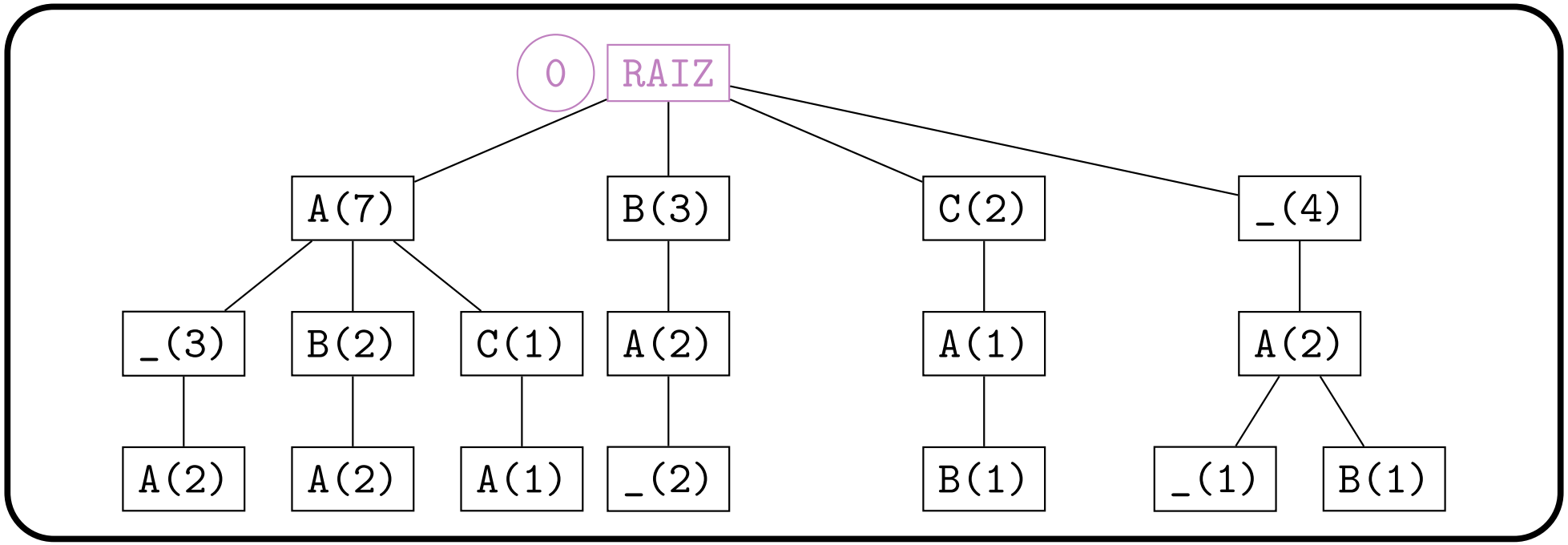


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

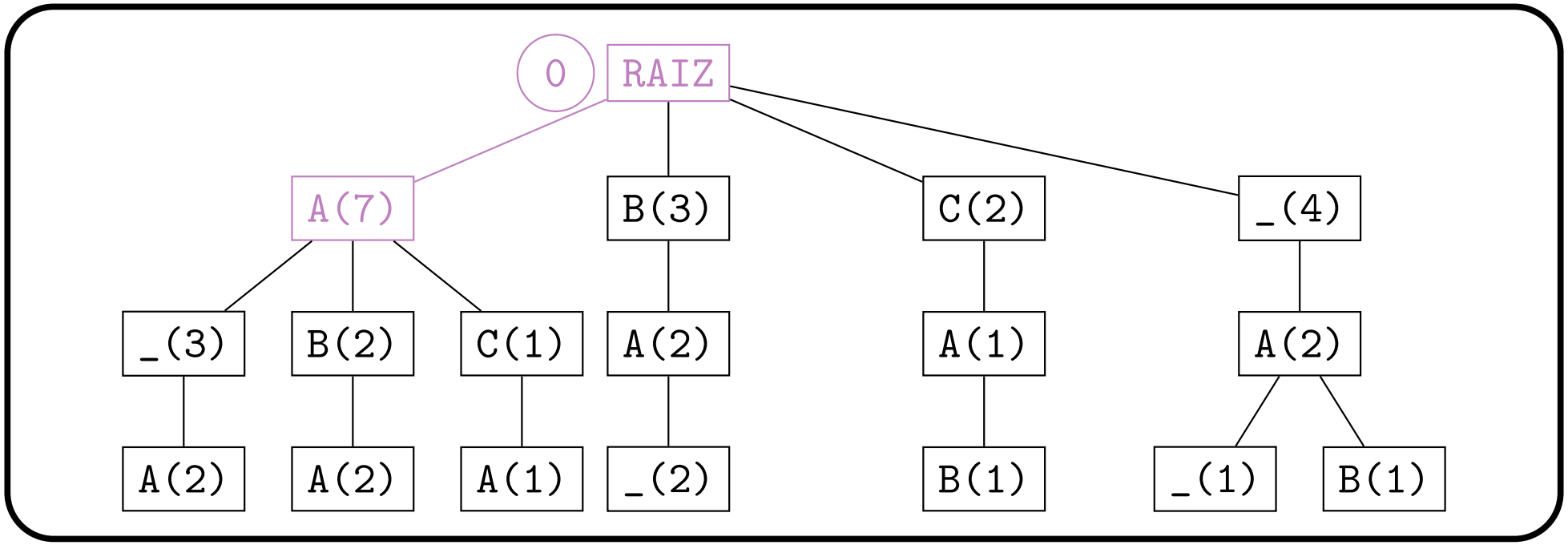


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

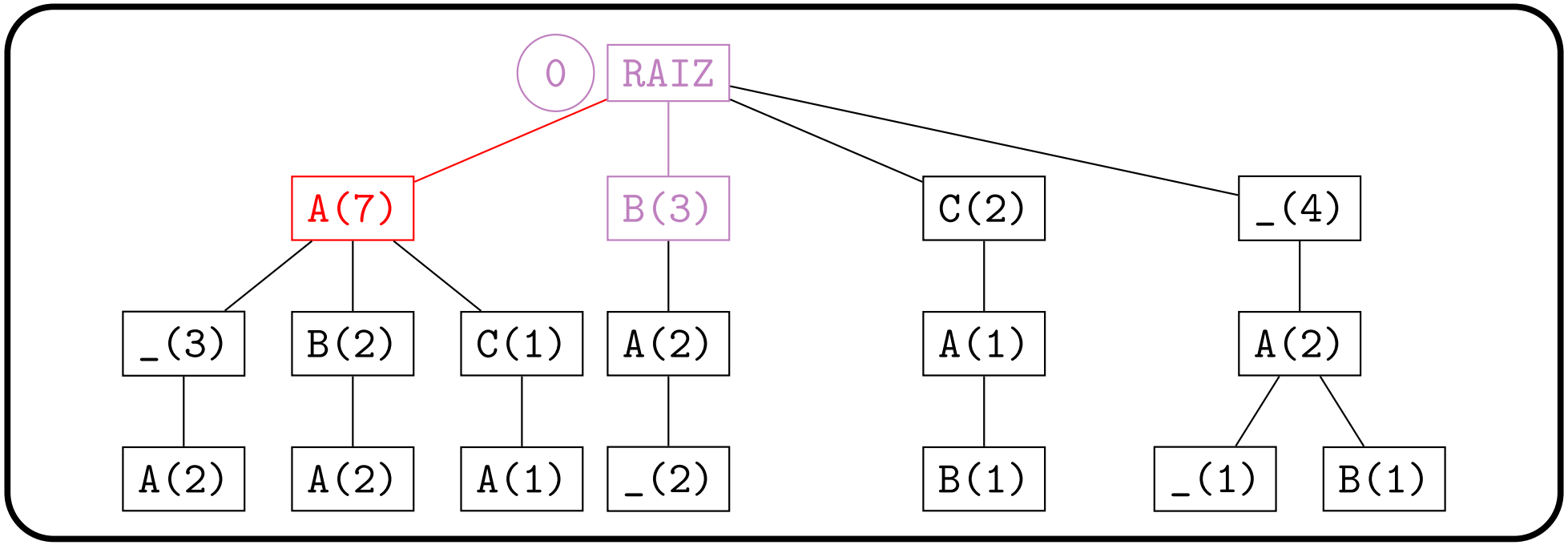


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

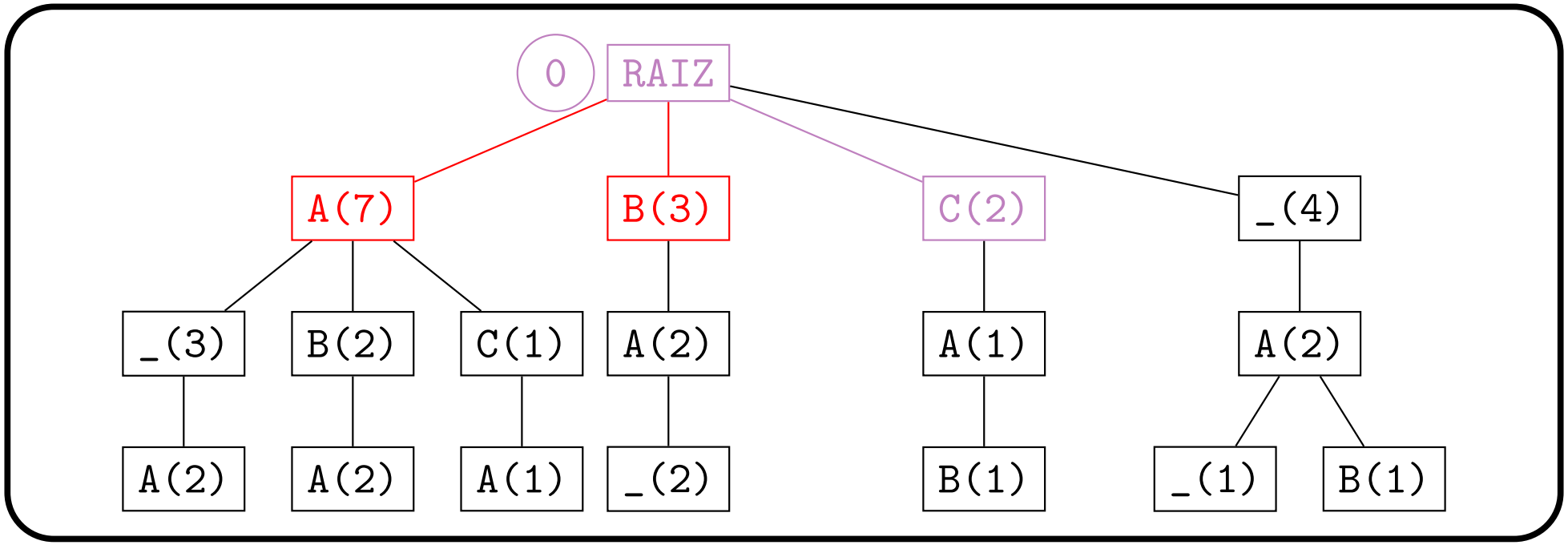


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto

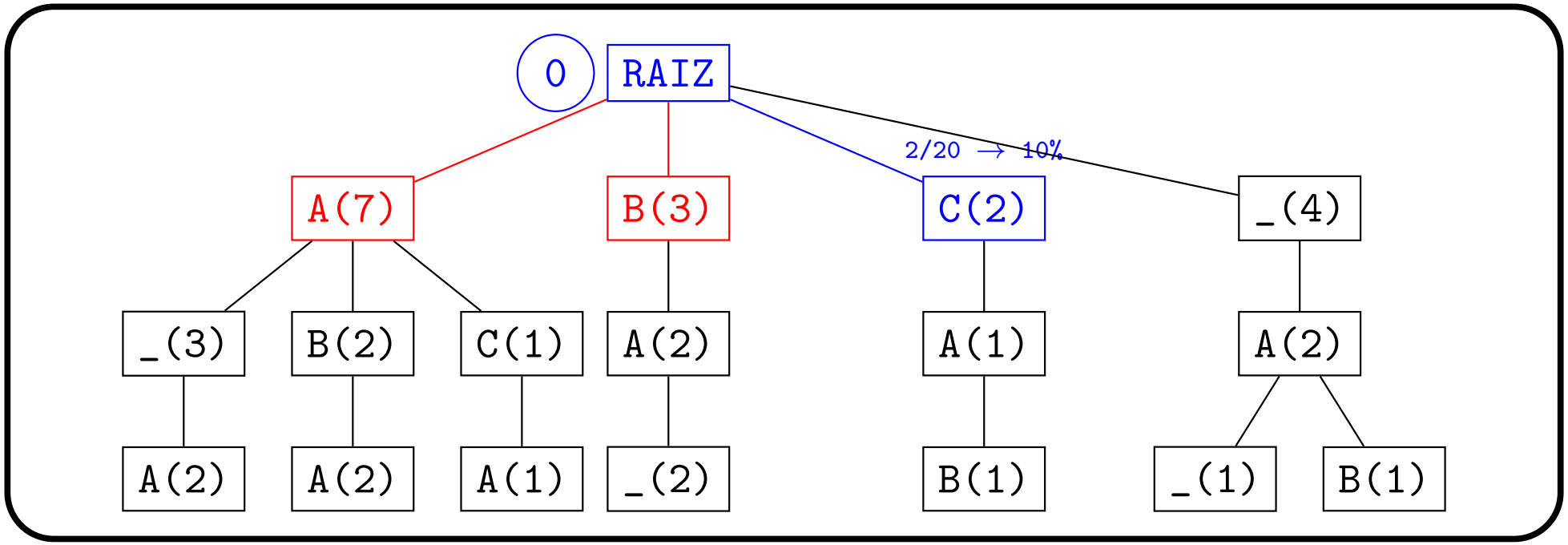


Metodologia do PPM

Exemplificação

A C A B A_ A _ A B A _ C

Contexto

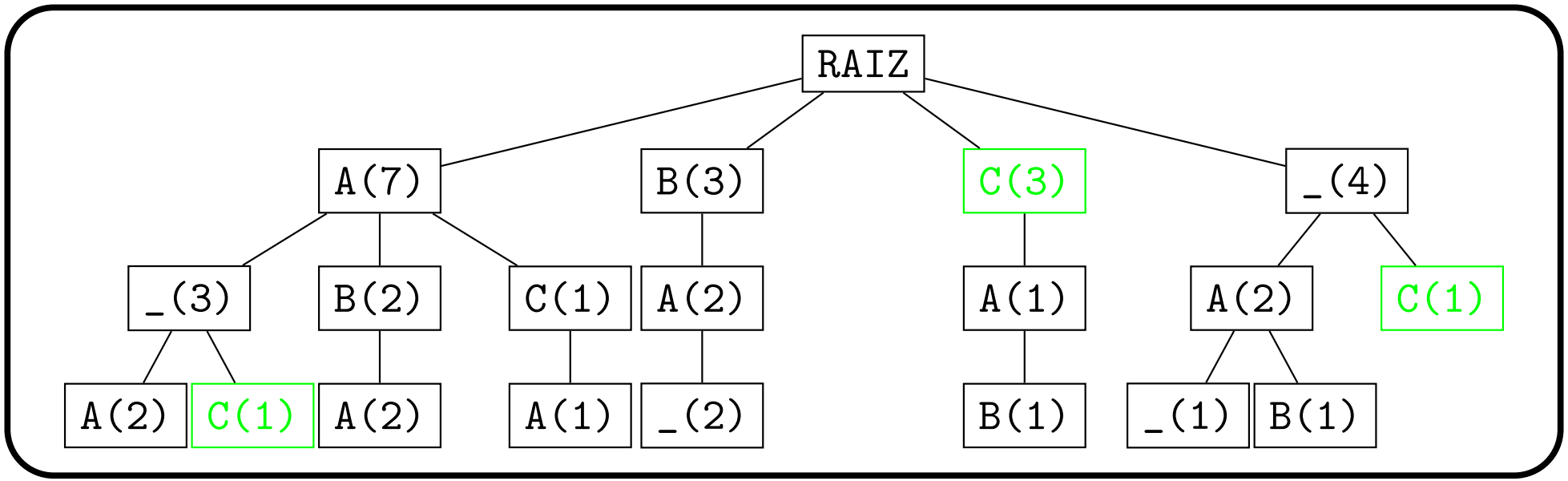


Metodologia do PPM

Exemplificação

A C A B A _ A _ A B A _ C

Contexto



Análise do PPM

Taxa de compressão

- Forte depêndencia da existência de padrões
- Sequências que custumam aparecer juntas utilizarão poucos bits
- Situação comum em dados textuais

Desempenho e uso de memória

- Quanto maior o número de padrões, mais rápida a compressão e descompressão
- Quanto maior o número de padrões, menor o custo em memória

Algoritmos avaliados

Algoritmo	**Implementação**	**Programa**
BWT	C	BZip2
Cod. Aritimética	C	Próprio
LZ	C	GZip
PPM	C	Próprio

Dados avaliados

TPC-H

- Benchmark para avaliação de performance em transações de BDs
- Banco de dados gerado de 1GB
- Adaptação do TPC-H para geração de dados orientados a coluna

Calgary corpus

- Benchmark para avaliação de compressores de dados

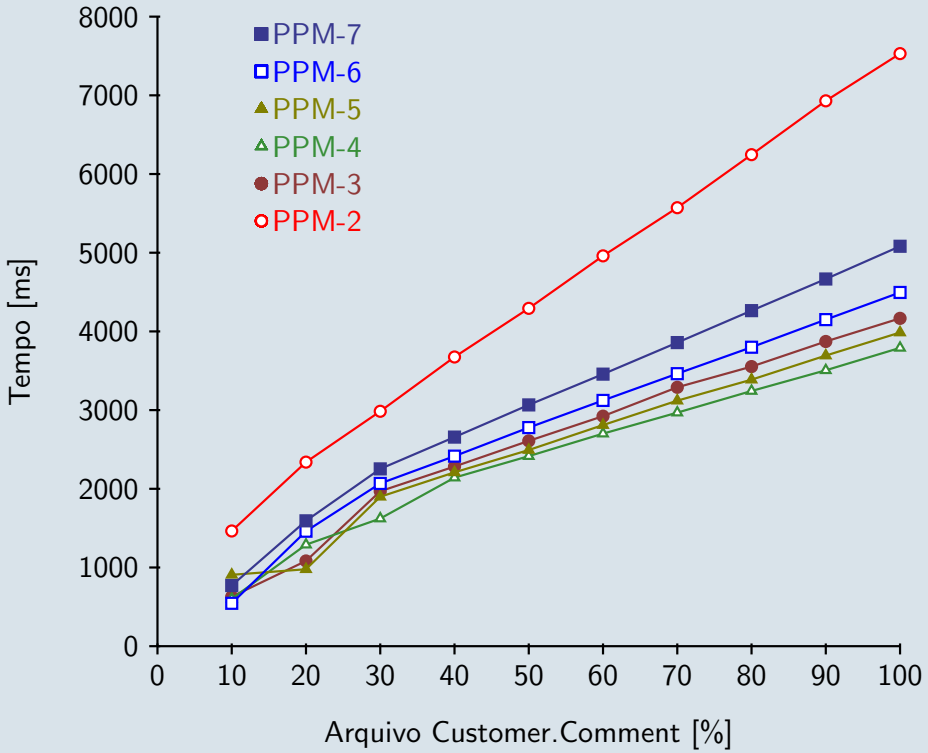
Determinação do tamanho de contextos

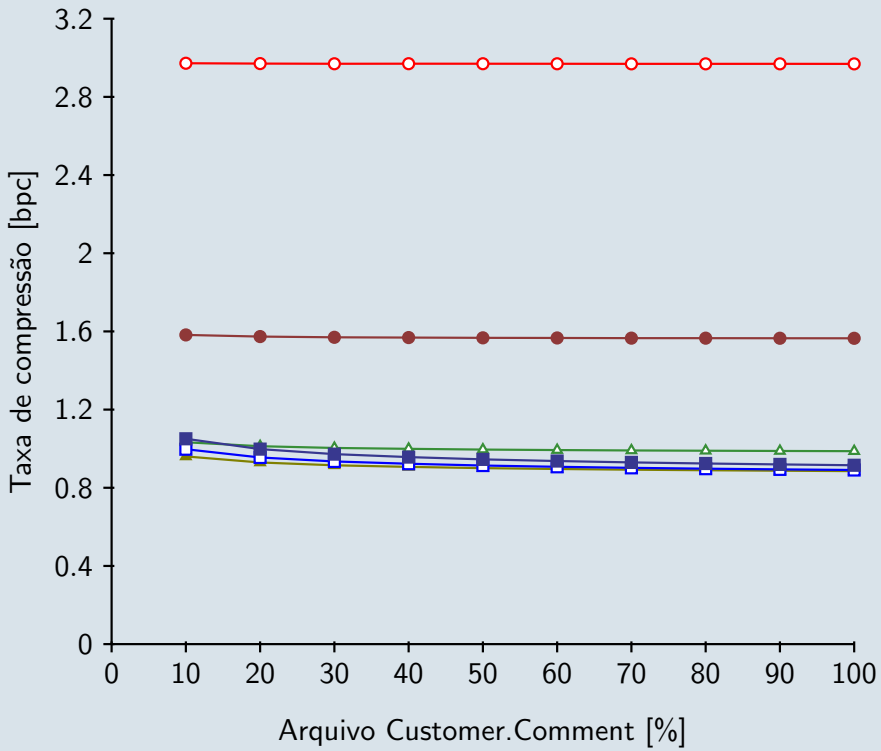
Objetivo

Determinar o melhor tamanho de contexto para compressão de dados orientados a coluna.

Determinação do tamanho de contextos

Resultado





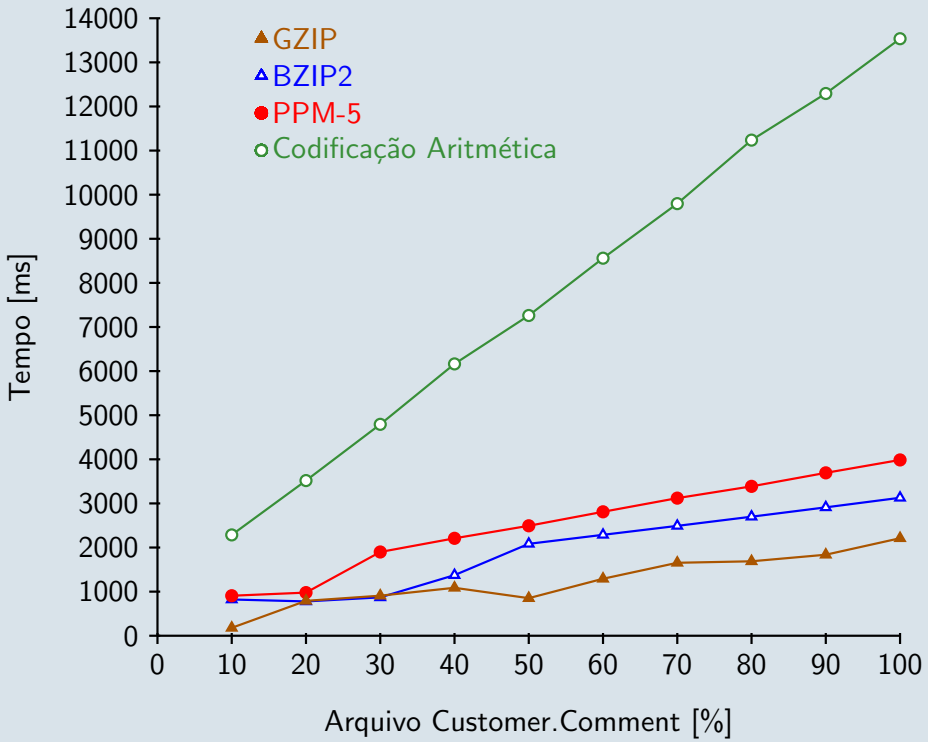
Análise dos algoritmos sob dados orientados a coluna

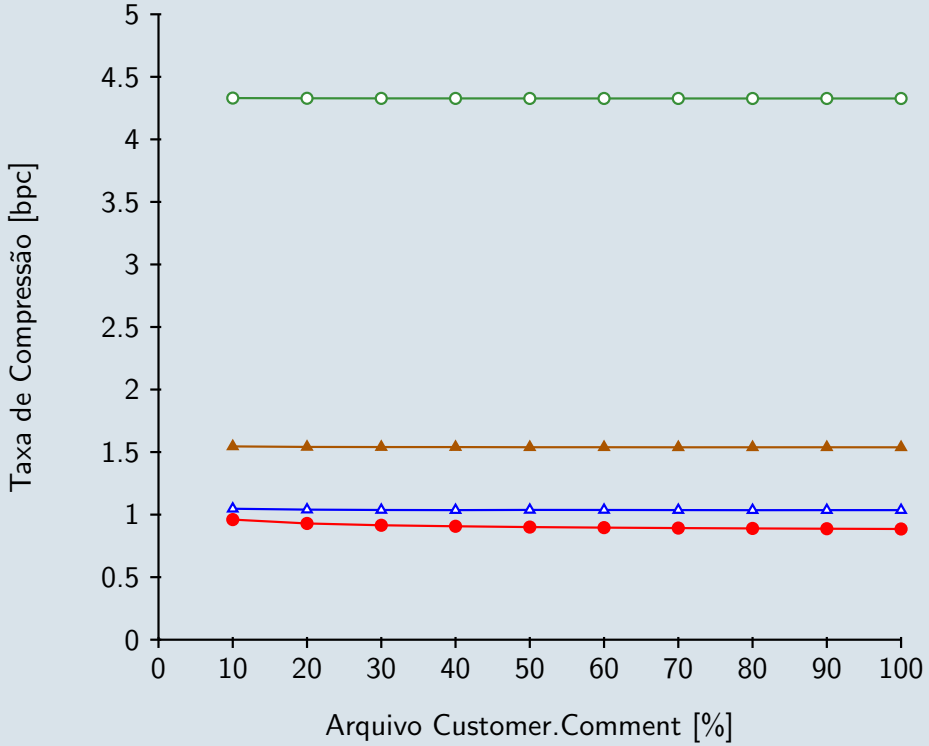
Objetivo

Determinar o comportamento dos algoritmos propostos anteriormente com arquivos orientados a coluna.

Análise dos algoritmos sob dados orientados a coluna

Resultado





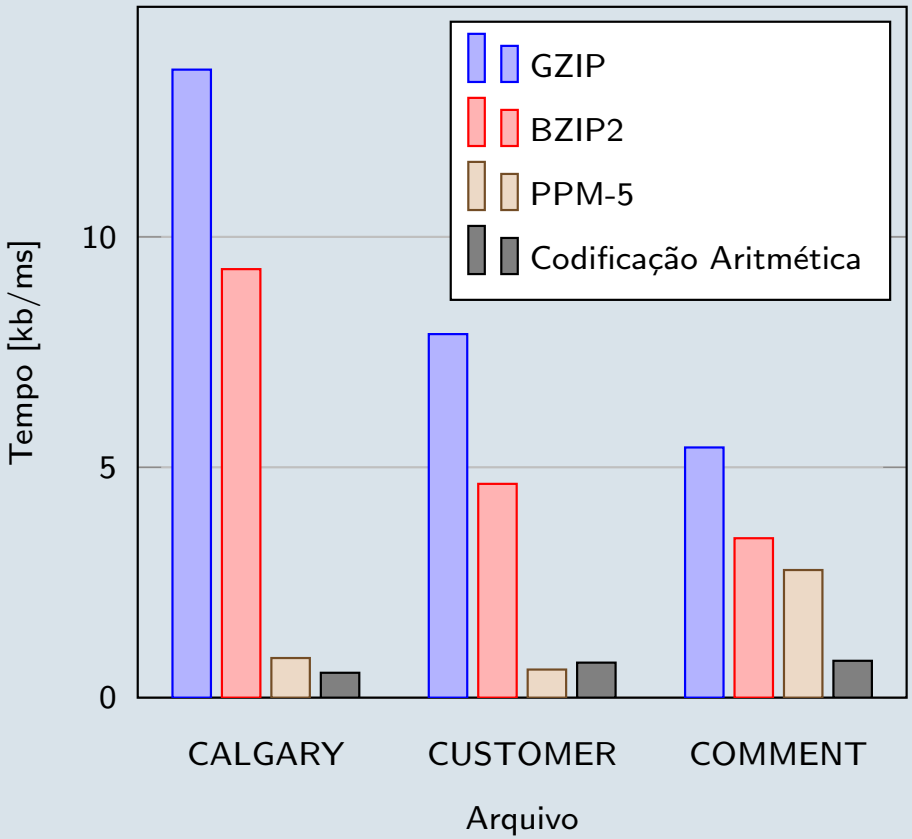
Análise dos algoritmos sob dados variados

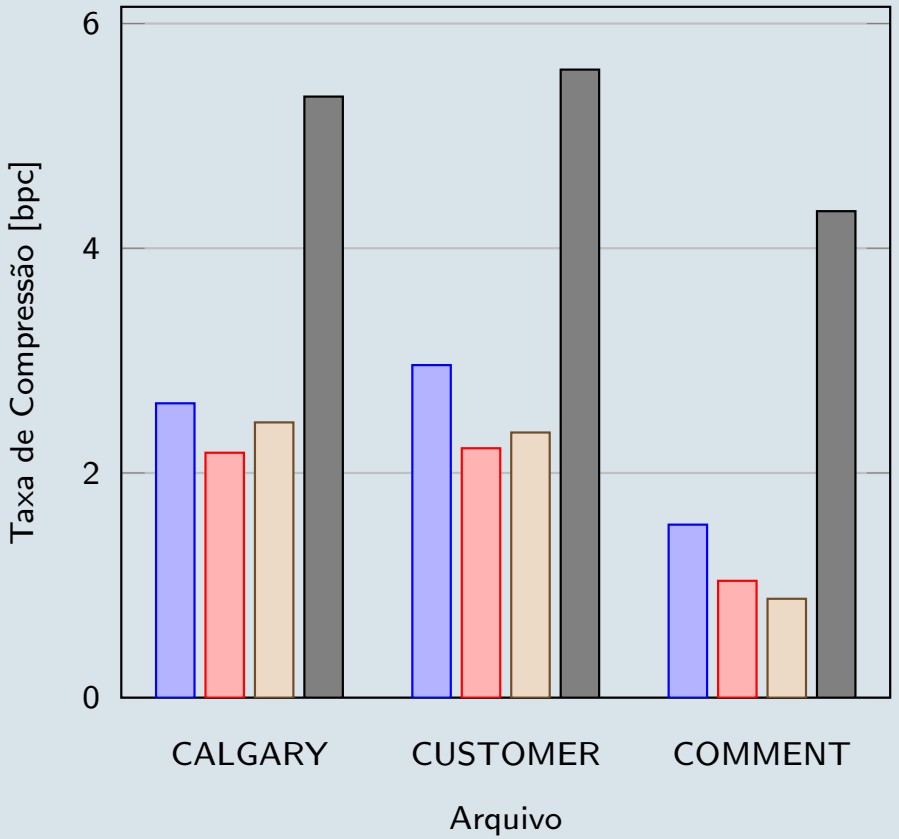
Objetivo

Determinar o comportamento dos algoritmos propostos anteriormente com arquivos variados presentes no corpus Calgary e em dados orientados a registros.

Análise dos algoritmos sob dados variados

Resultado





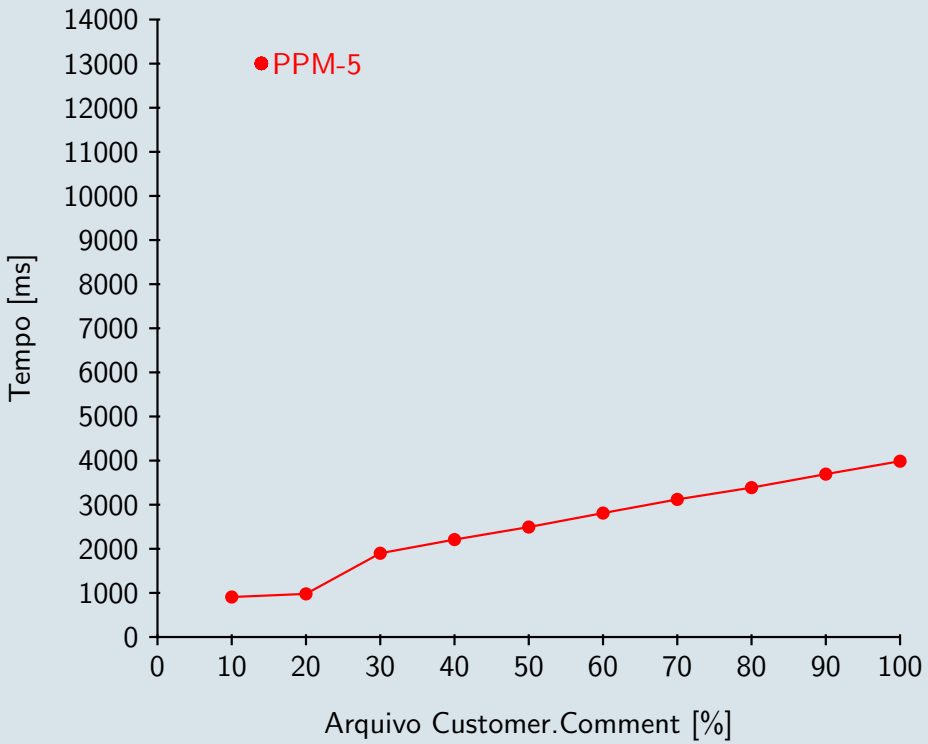
Conclusões

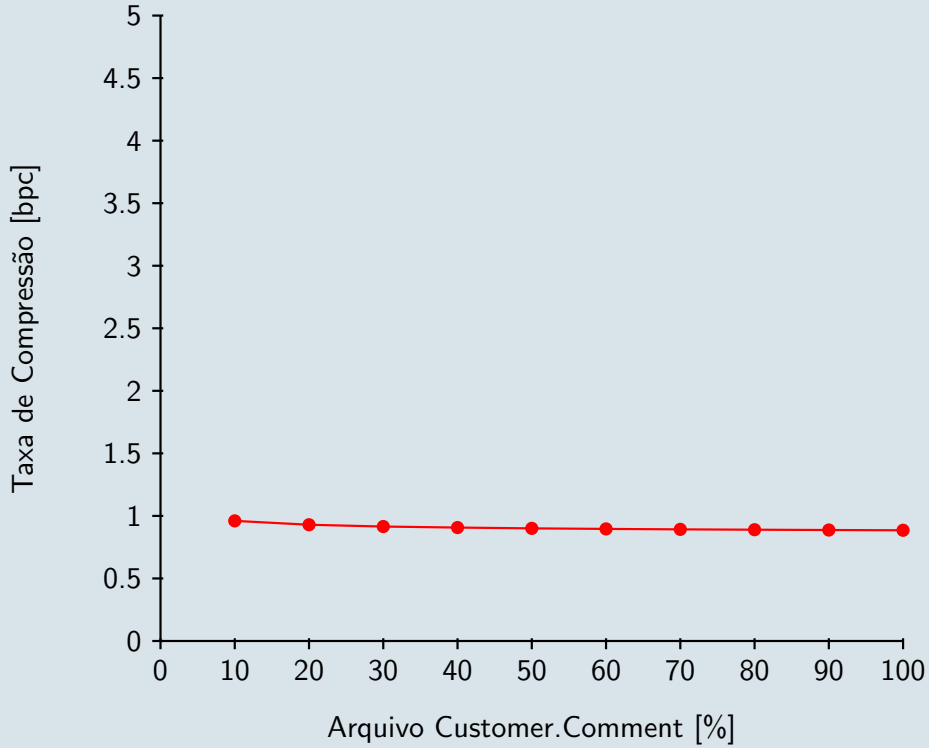
- Arquivos orientados a coluna são bem explorados por compressores baseados em padrões
- O método PPM mostrou um custo benefício aceitável na relação compressão/desempenho
- O método PPM mostra-se com aplicação viável sob dados orientados a coluna

Trabalhos futuros

Atualização da árvore de contexto

- Analisar métricas de compressão (bpc) durante a execução do algoritmo
- Determinar pontos de não atualização da árvore
- Otimizar o tempo de execução do método PPM





Compressão de Arquivos Orientados a Coluna com PPM

Perguntas?

Universidade Federal de Santa Maria
Santa Maria - RS

Vinícius Fülber Garcia
vfulber@inf.ufsm.br

Sérgio L. S. Mergen
mergen@inf.ufsm.br