Detecção de Réplicas de Sítios Web em Máquinas de Busca Usando Aprendizado de Máquina

Rickson Guidolini

Orientador: Nivio Ziviani

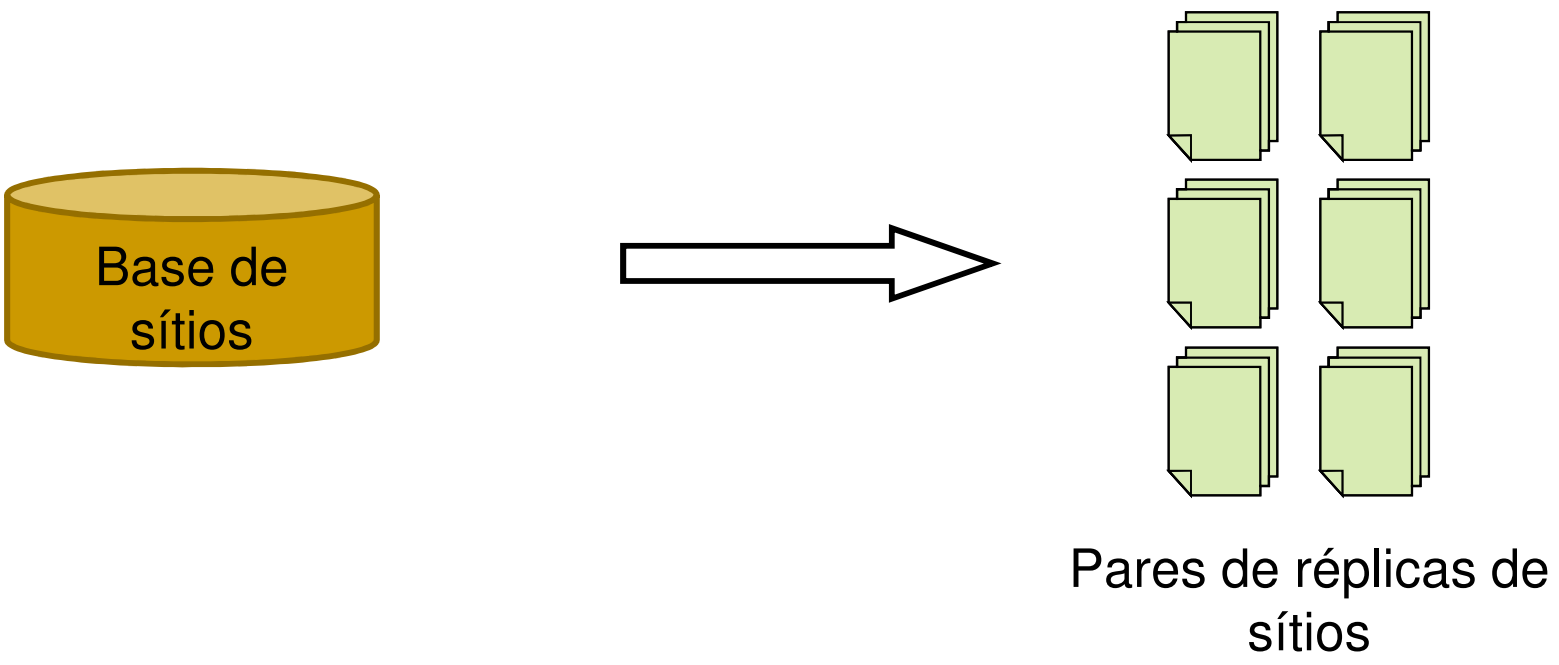
Co-orientador: Adriano Veloso

Universidade Federal de Minas Gerais

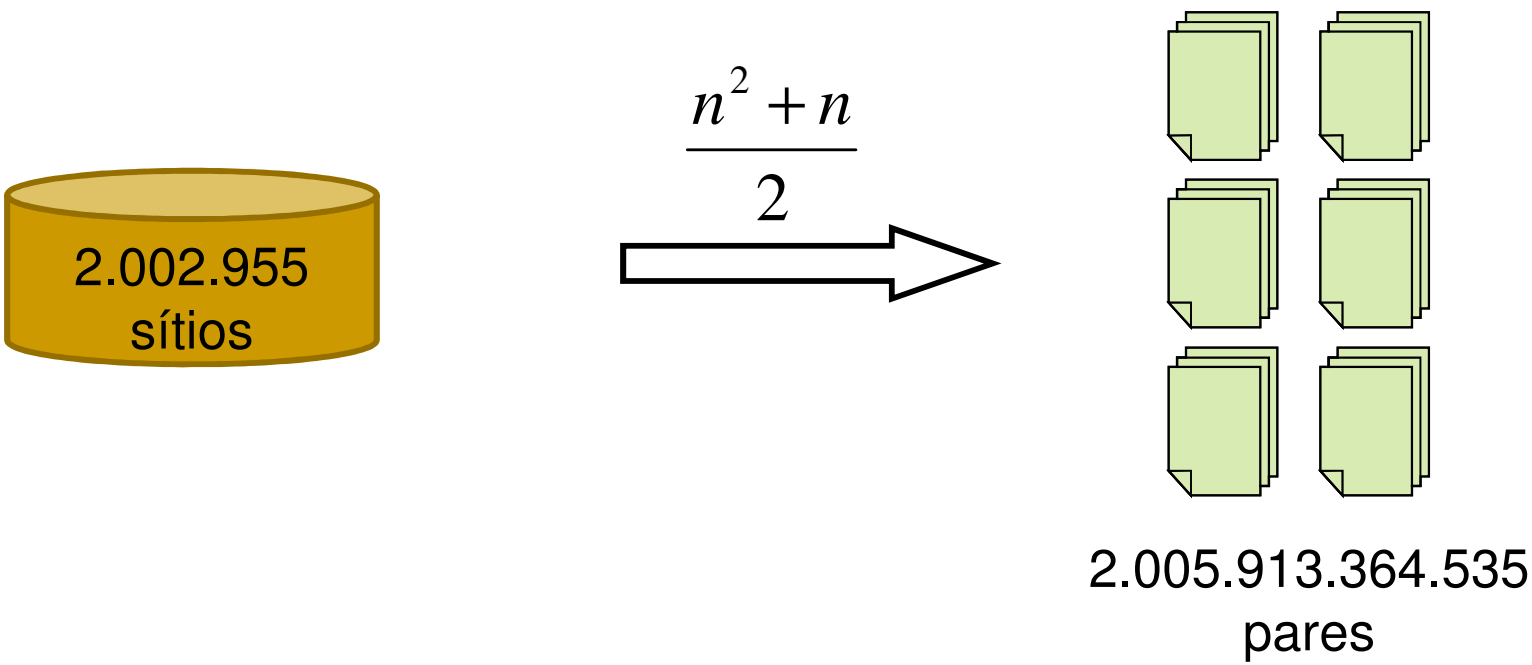
Agenda

- Introdução
  - Definição do Problema
  - Desafios
  - Trabalhos relacionados
- Algoritmo Proposto (DREAM)
  - Características utilizadas
  - As três fases do algoritmo DREAM
- Resultados
  - Coleção de dados
  - Comparação entre os algoritmos
- Conclusões e Trabalhos Futuros
- Contribuições

O Problema Abordado



Algoritmo Ingênuo

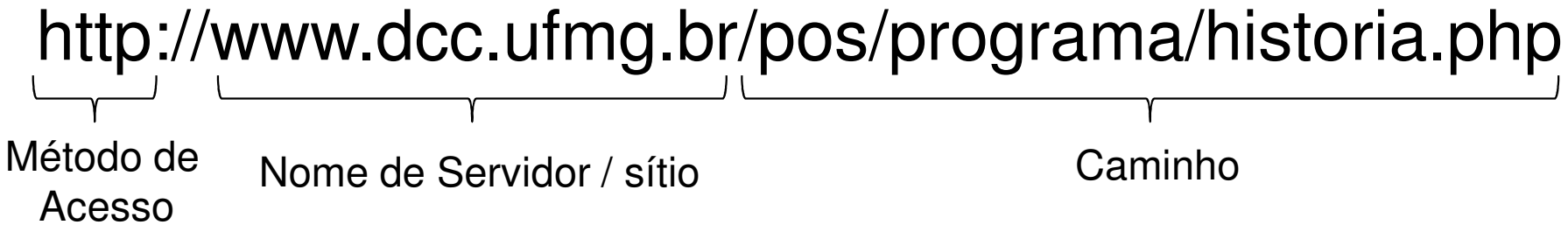


Definição do Problema

- Dois sítios A e B são ditos serem réplicas se:
  I. Uma alta porcentagem dos caminhos de A são válidos em B e vice-versa
  II. Esses caminhos comuns designam documentos com conteúdo similar entre os dois sítios

Definição do Problema

- Dois sítios A e B são ditos serem réplicas se:
  I. Uma alta porcentagem dos caminhos de A são válidos em B e vice-versa
  II. Esses caminhos comuns designam documentos com conteúdo similar entre os dois sítios



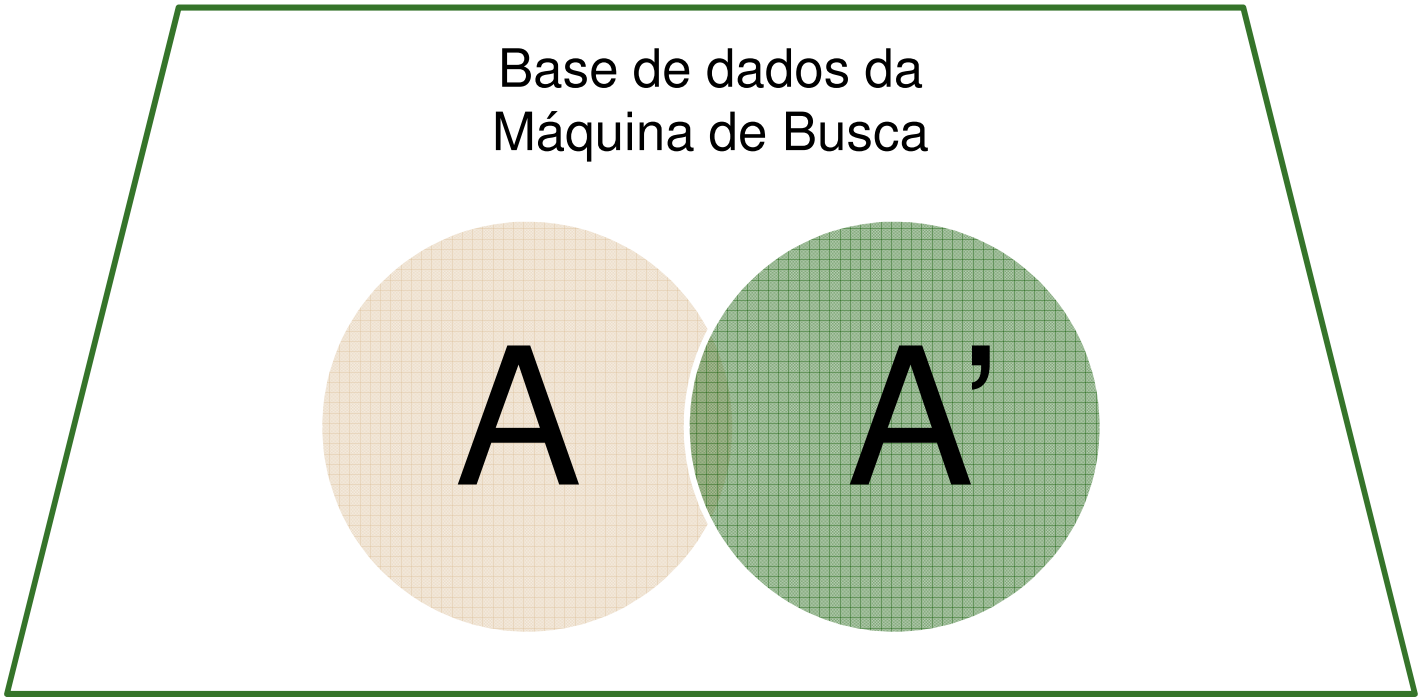
Custos Envolvidos

- Verificação de replicação implica em
  - Coleta de páginas web
  - Comparação entre conteúdos textuais
- Para uma amostragem de 10 caminhos:

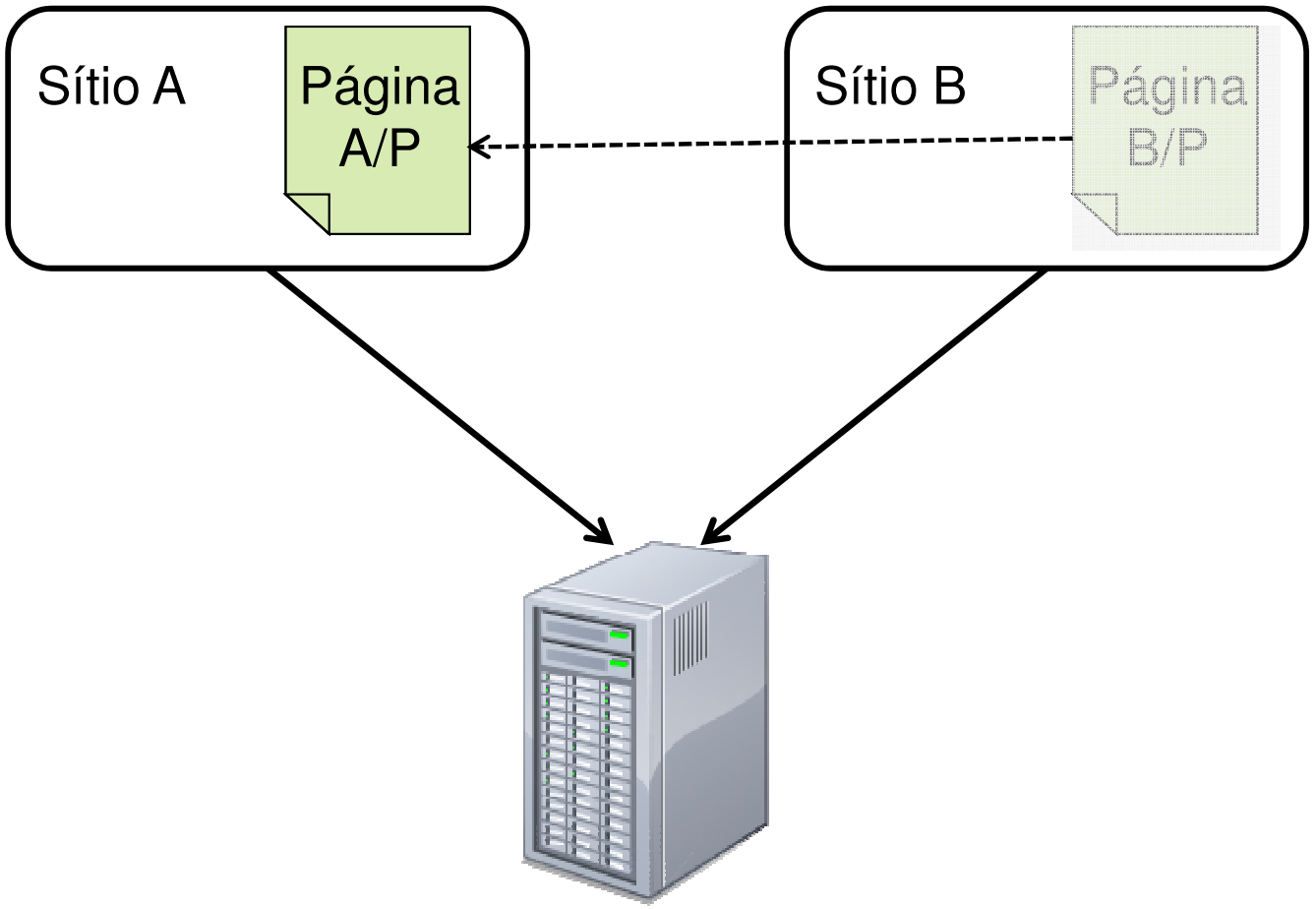
$$downloads = \frac{n^2 + n}{2} \times 4 \times |\, amostra \,| = 80.236.534.581.400$$

$$comparações = \frac{n^2 + n}{2} \times |\, amostra \,| = 20.059.133.645.350$$

Desafios: Cobertura por Sítios



Desafios: Caminhos Trocados

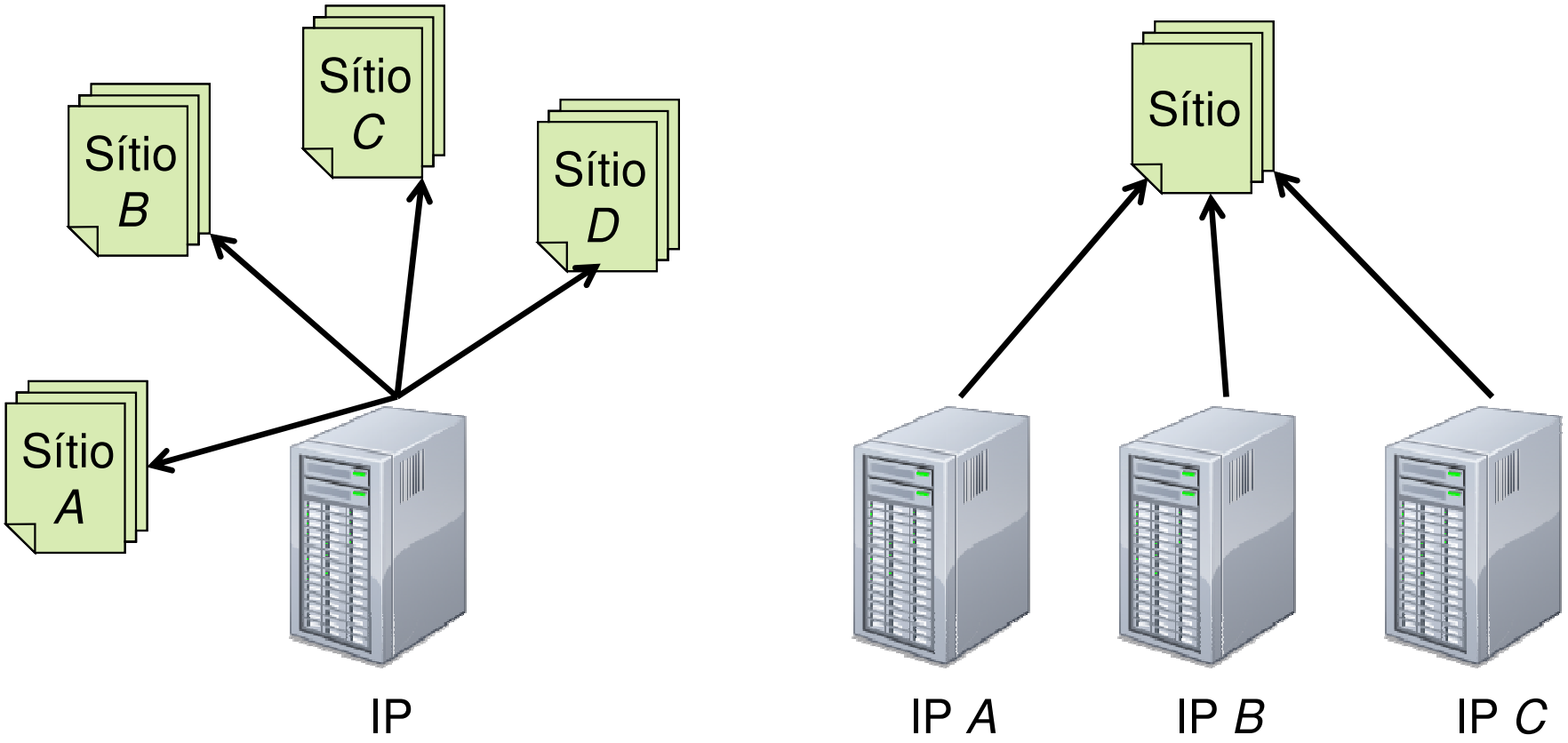


Desafios: Conteúdo Volátil

Desafios: Conteúdo Regional

Desafios: Conteúdo Regional

Desafios: Relação IP-Sítio é N:N



As Principais Causas da Replicação de Sítios Web

- Múltiplos nomes de domínio
  - Banco do Brasil
- Balanceamento de carga
  - Filiais
- Franquia
  - www.abril.com.br e www.abril.uol.com.br
- Razões Sociais
  - *Protein Data Bank*

Importância da Detecção de Réplicas

- Economia de recursos
  - Armazenamento, processamento, consumo de banda
- Previne anomalias no *ranking*
  - Informações de conectividade também são replicadas
- Evita repetições nas respostas finais
  - Respostas repetidas aborrecem os usuários
- Aumenta a diversidade dos sítios cobertos
  - Novos sítios podem ser coletado no lugar dos sítios replicados

Objetivo

- Avaliar abordagens baseadas em aprendizado de máquina para o problema de detecção de réplicas de sítios web
  - É difícil encontrar o valor ideal das características
  - Várias características podem ser combinadas
- Resultado: DREAM – Detecção de RÉplicas usando Aprendizado de Máquina

Trabalhos Relacionados

- Bharat & Broder (1999)
  - Estudo sobre replicação na Web
- Bharat, Broder, Dean, & Henzinger (2000)
  - Vários métodos para detecção de réplicas de sítios
- Cho, Shivakumar, & Garcia-Molina (2000)
  - Coleções replicadas
- da Costa Carvalho, de Moura, da Silva, Berlt & Bezerra (2007)
  - Uso eficiente de informações a cerca do conteúdo
  - NormPaths: Estado-da-arte e *baseline*

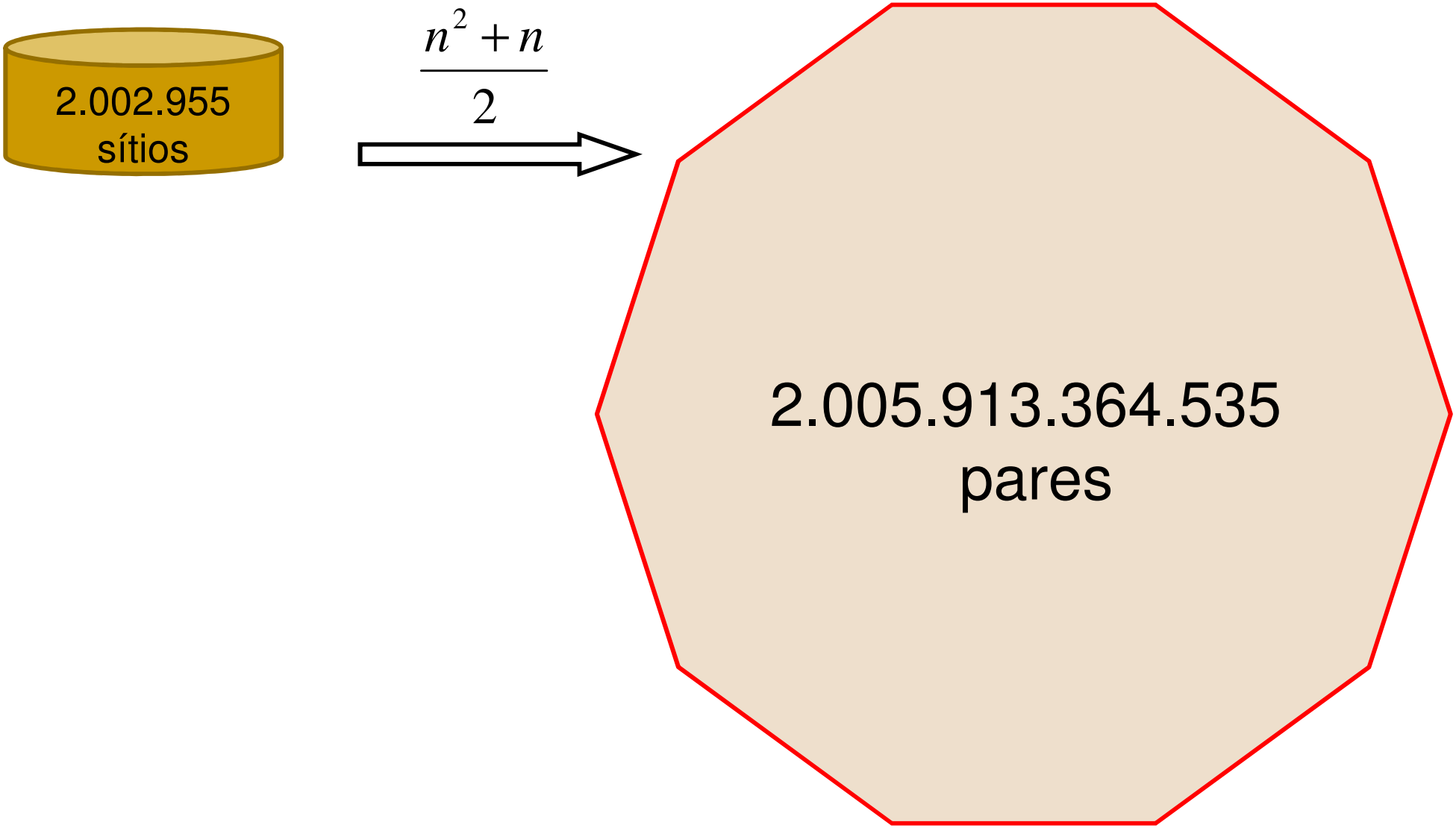
O Algoritmo DREAM

- Fase 1: Características gerais são utilizadas para selecionar um conjunto de pares com algum potencial para serem réplicas
- Fase 2: Características específicas são utilizadas pelo modelo aprendido (aprendizado de máquina), obtendo uma lista refinada de pares de sítios
- Fase 3: Respostas do modelo de detecção de réplicas são validadas

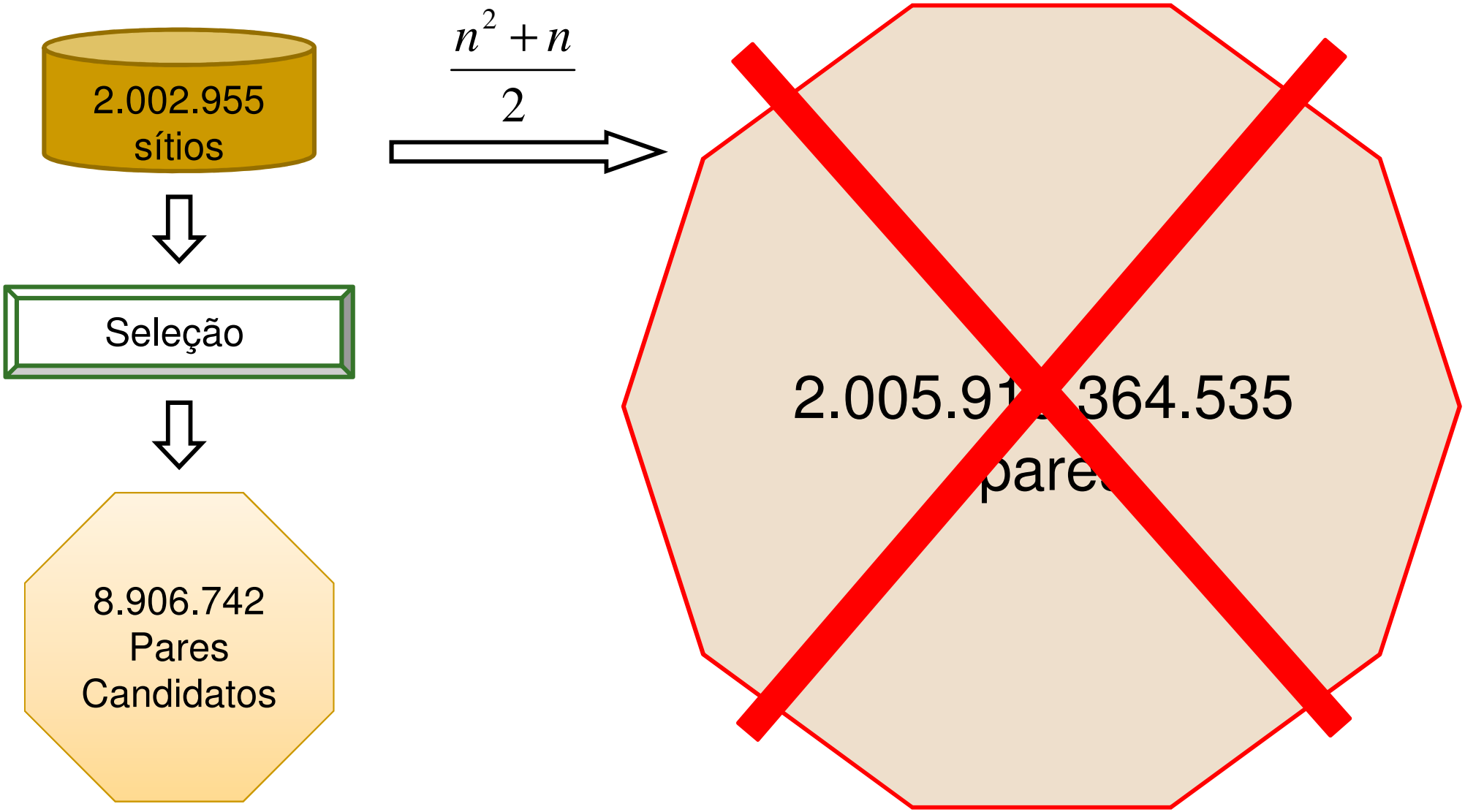
Características Utilizadas

Característica	Fase
Caminho da URL	1ª Fase
Assinatura do Conteúdo	1ª Fase
Distância de Edição	**2ª Fase**
Diferença de Segmentos	**2ª Fase**
Correlação entre nomes de Servidor	2ª Fase
Quatro Octetos	2ª Fase
Três Octetos	2ª Fase
Correlação entre caminhos completos	2ª Fase
Caminho e conteúdo	2ª Fase

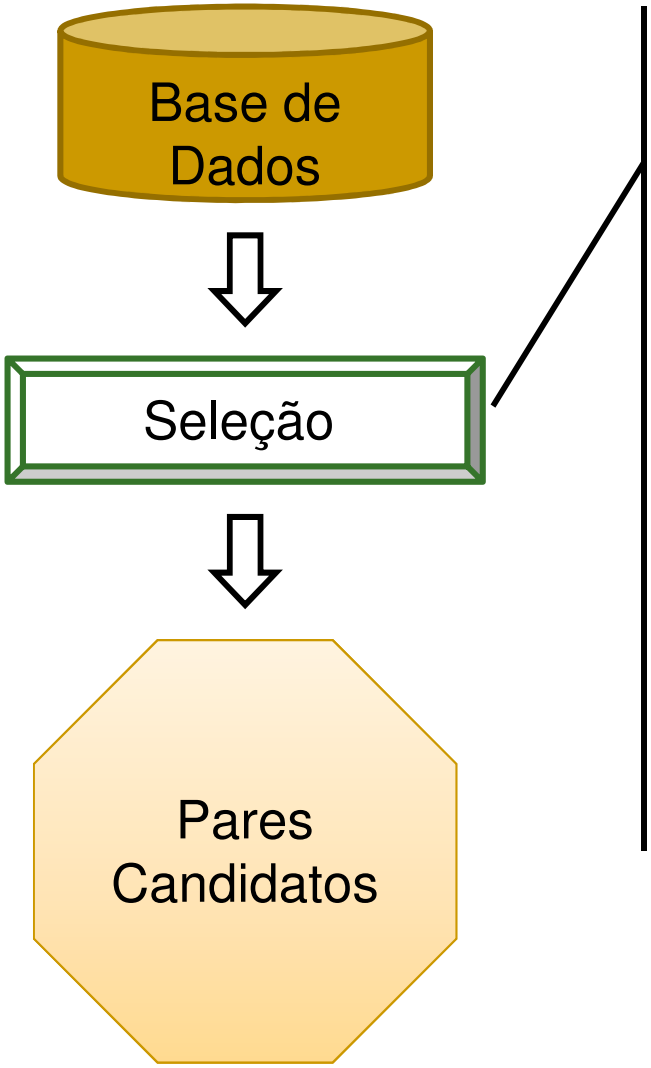
Algoritmo Ingênuo

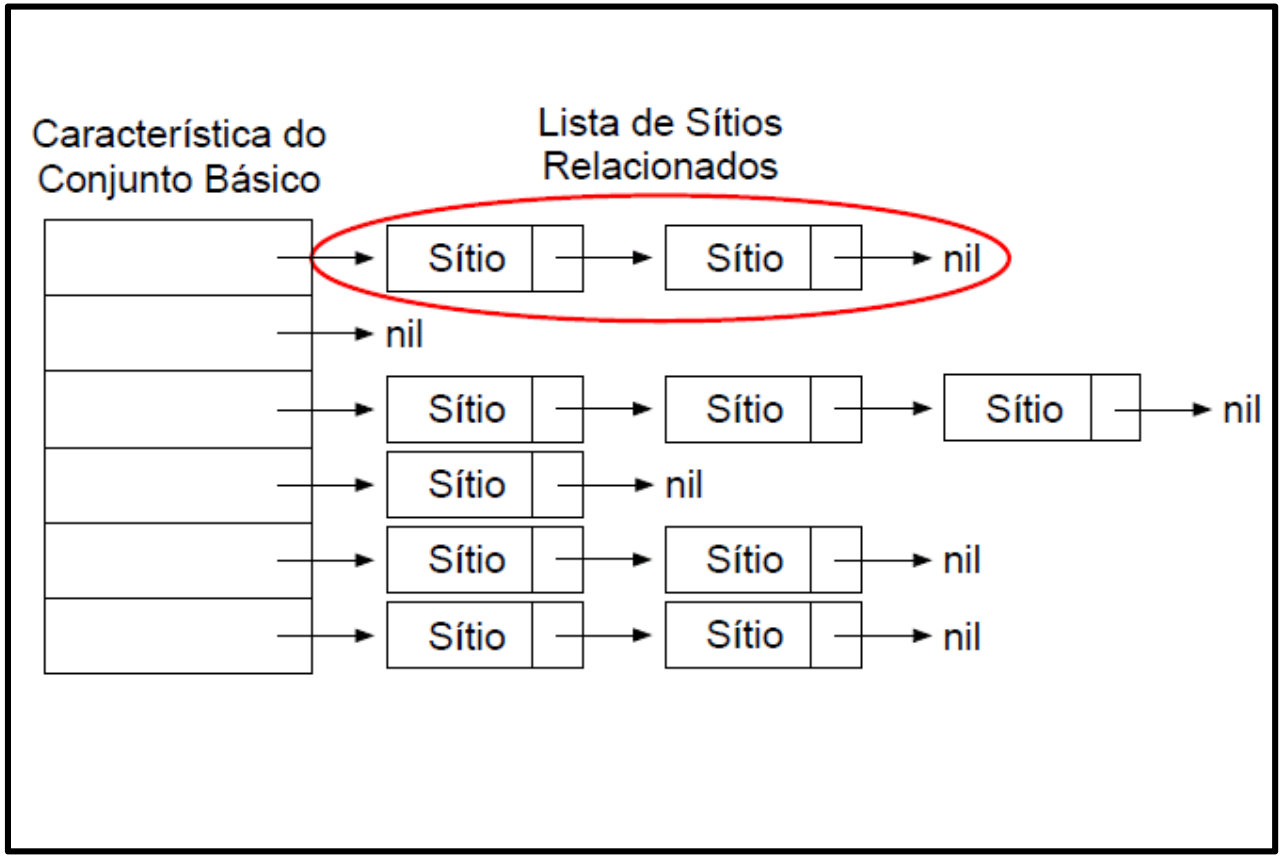


1ª Fase *versus* Algoritmo Ingênuo

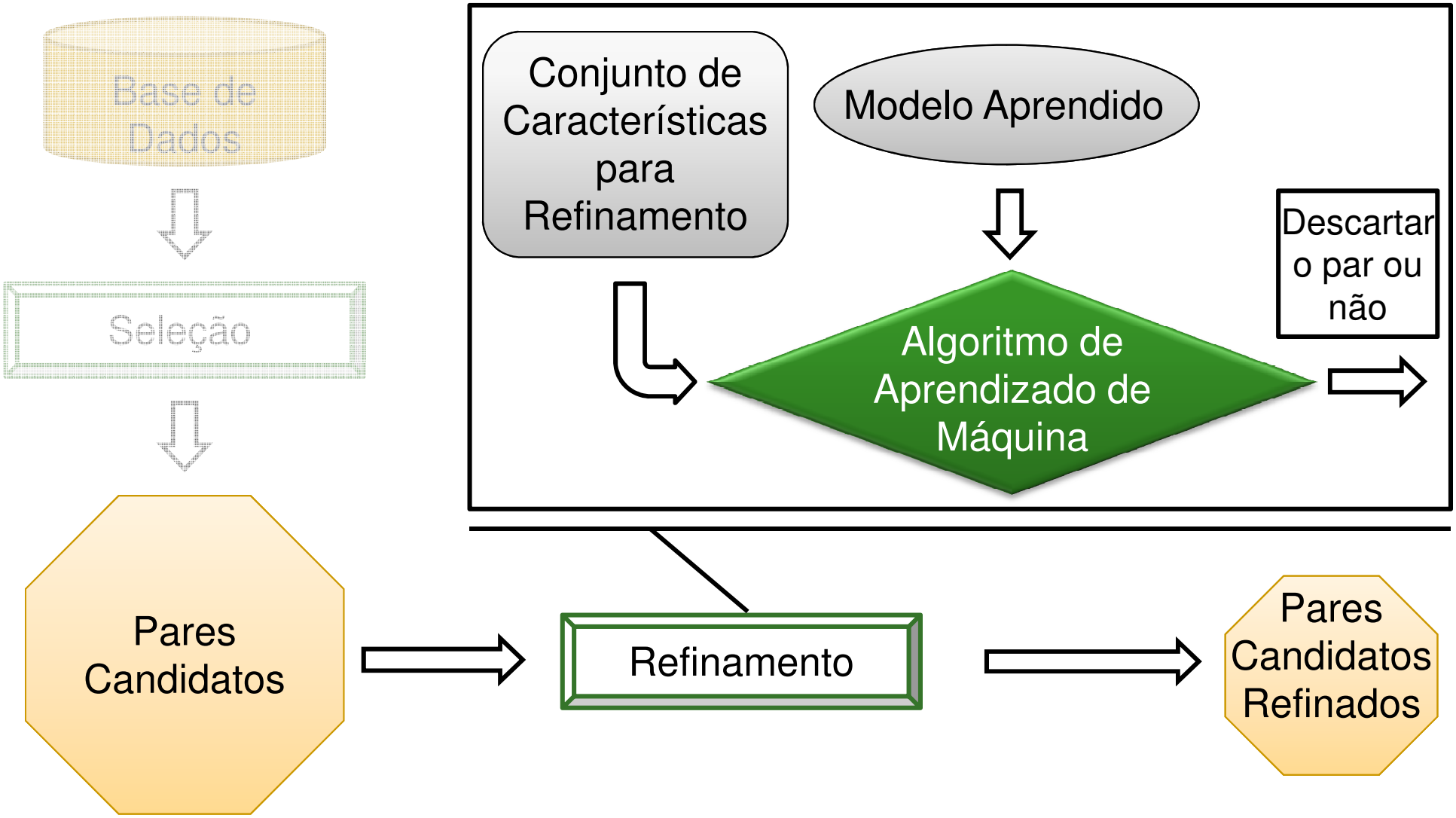


1ª Fase: Seleção de pares candidatos

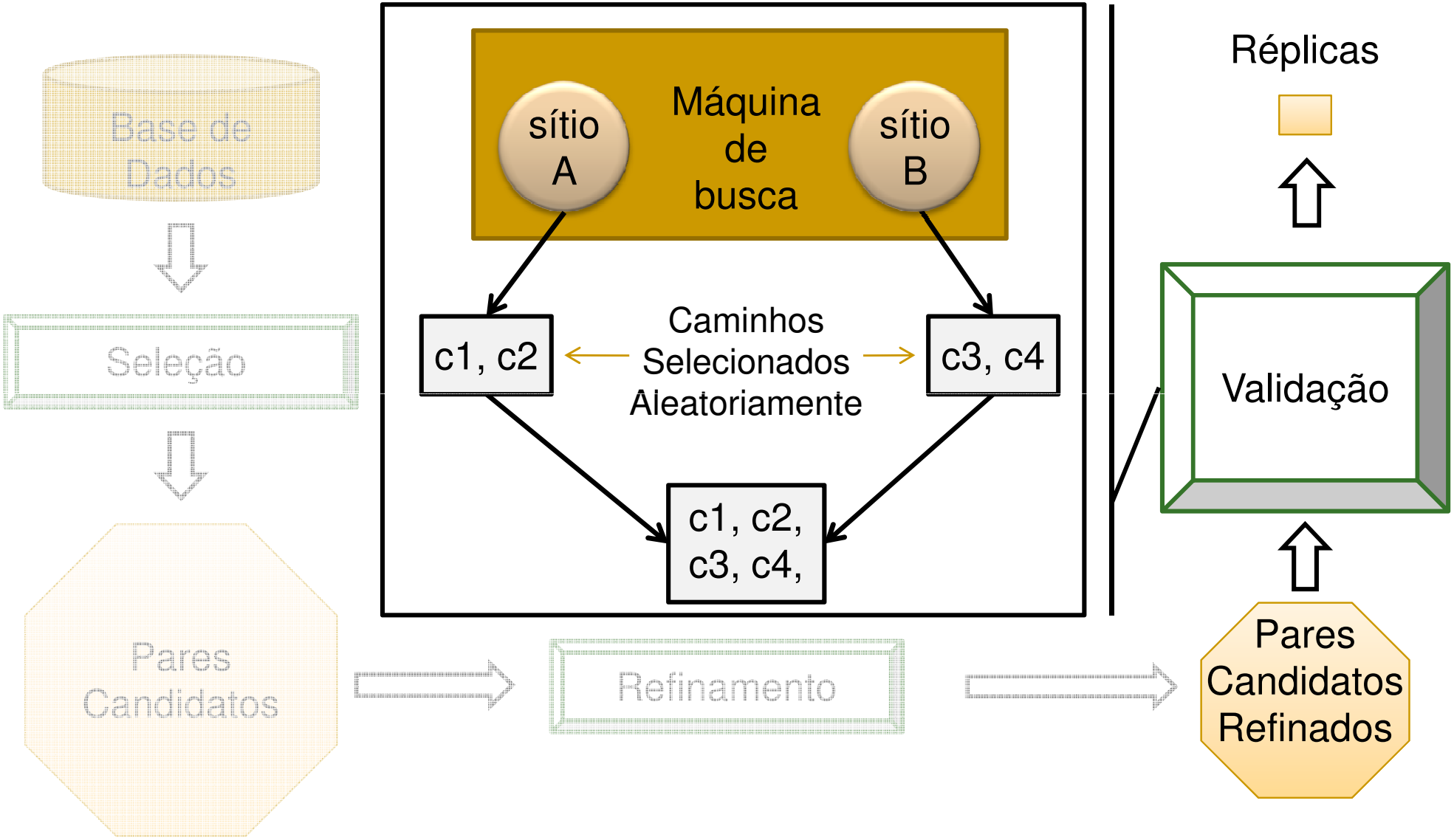




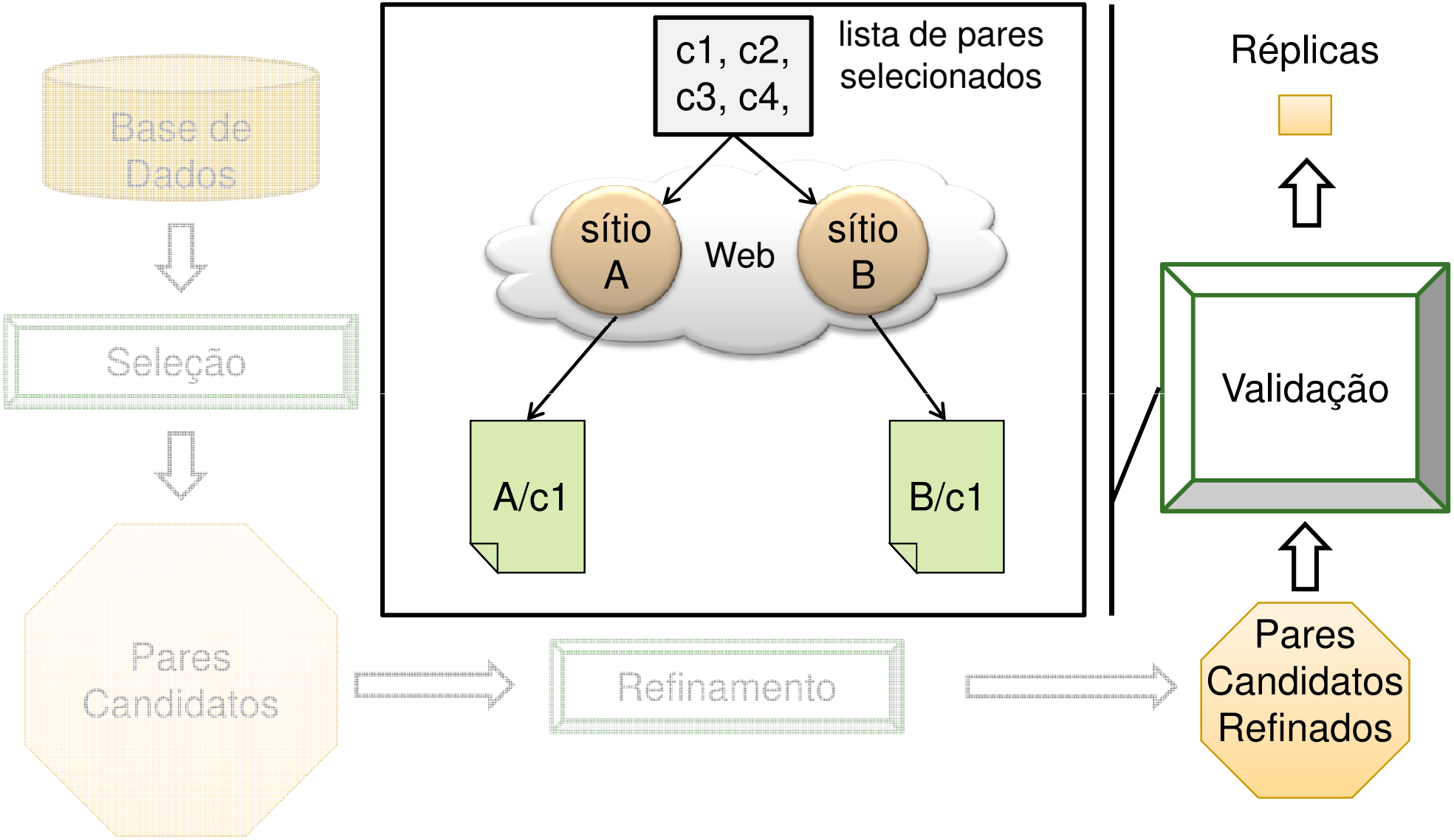
2ª Fase: Aplicação do Modelo Aprendido



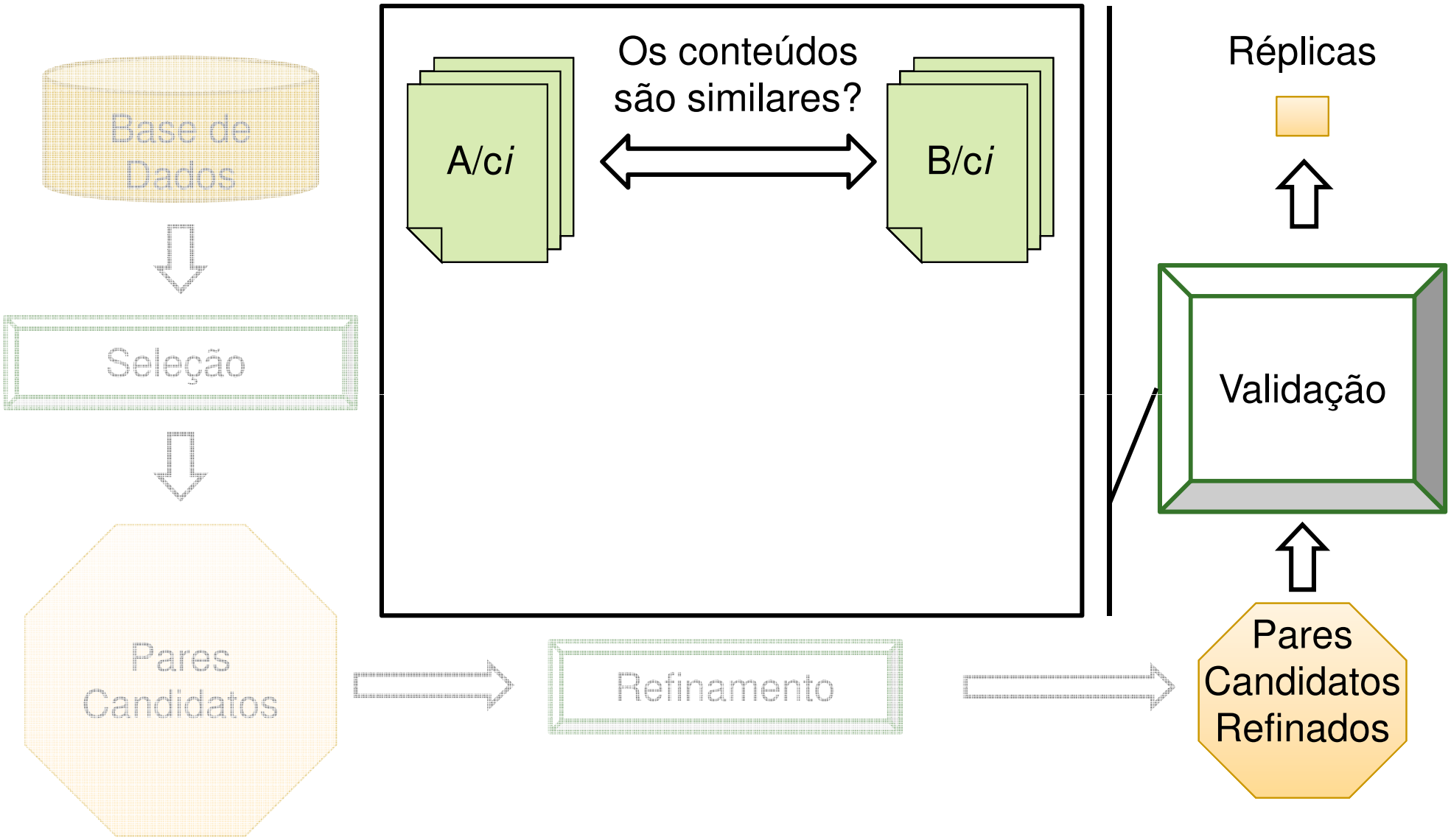
3ª Fase: Seleção Aleatória de Caminhos



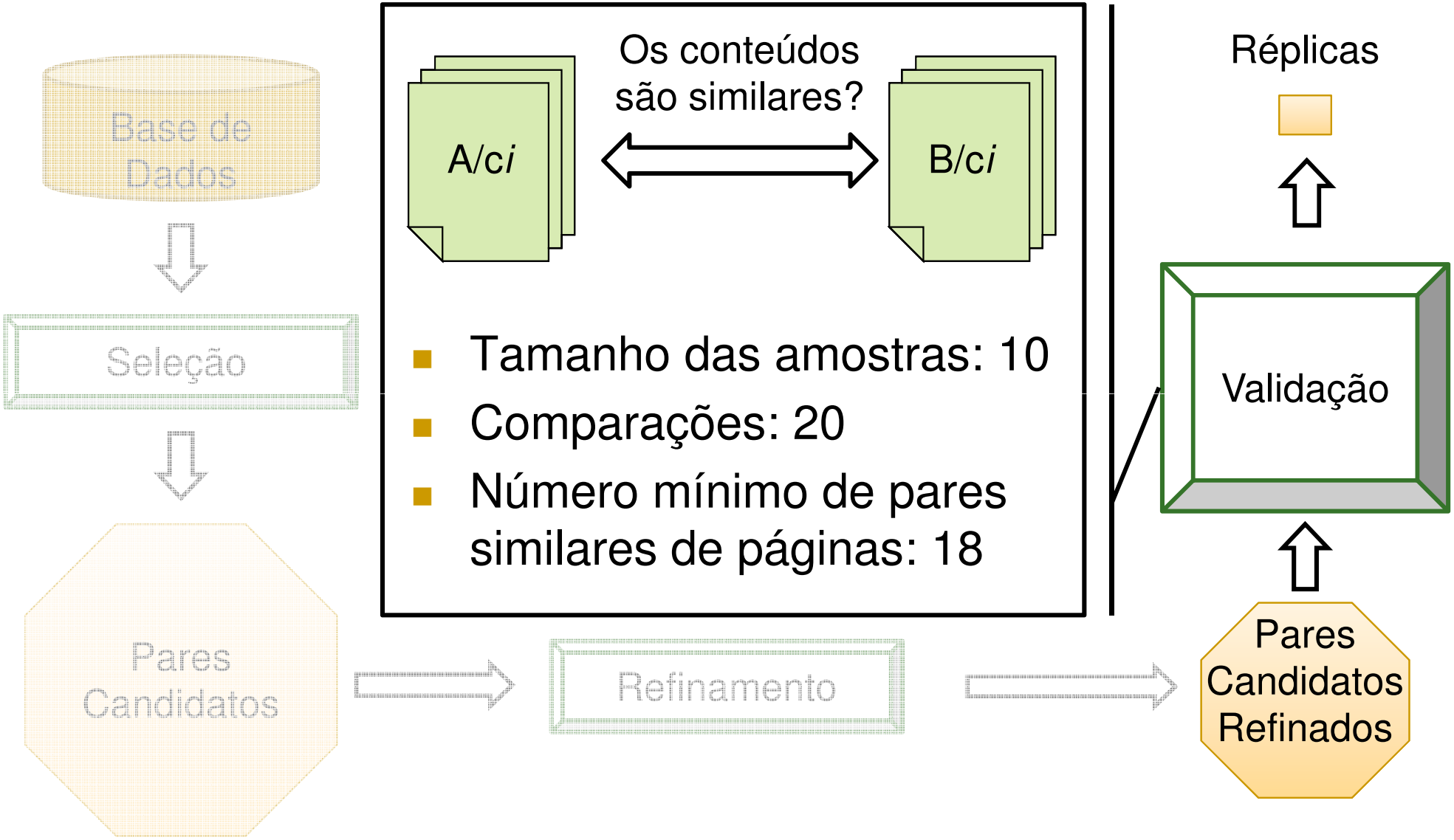
3ª Fase: Coleta da Lista de Caminhos



3ª Fase: Comparação Entre Conteúdos



3ª Fase: Comparação Entre Conteúdos



Algoritmos de Aprendizado de Máquina Estudados

- Árvore de decisão
  - DREAM-DT
- Combinação de árvores *(Random Forest)*
  - DREAM-RF
- Classificador associativo (LAC)
  - DREAM-LAC
- Máquina de vetor de suporte (SVM)
  - DREAM-SVM

Geração do Gabarito

- WBR2010B: **2.002.955** sítios conhecidos
- Total de pares possíveis: **2.005.913.364.535**
- Custos da verificação automática torna a abordagem ingênua proibitiva
- Como nosso objetivo não é encontrar todas as réplicas da WBR2010B
  - Redução da dimensionalidade (1ª fase DREAM)
- O desempenho do *Baseline* não foi afetado
  - Recupera um subconjunto do conjunto selecionado pela 1ª fase do DREAM

Geração do Gabarito

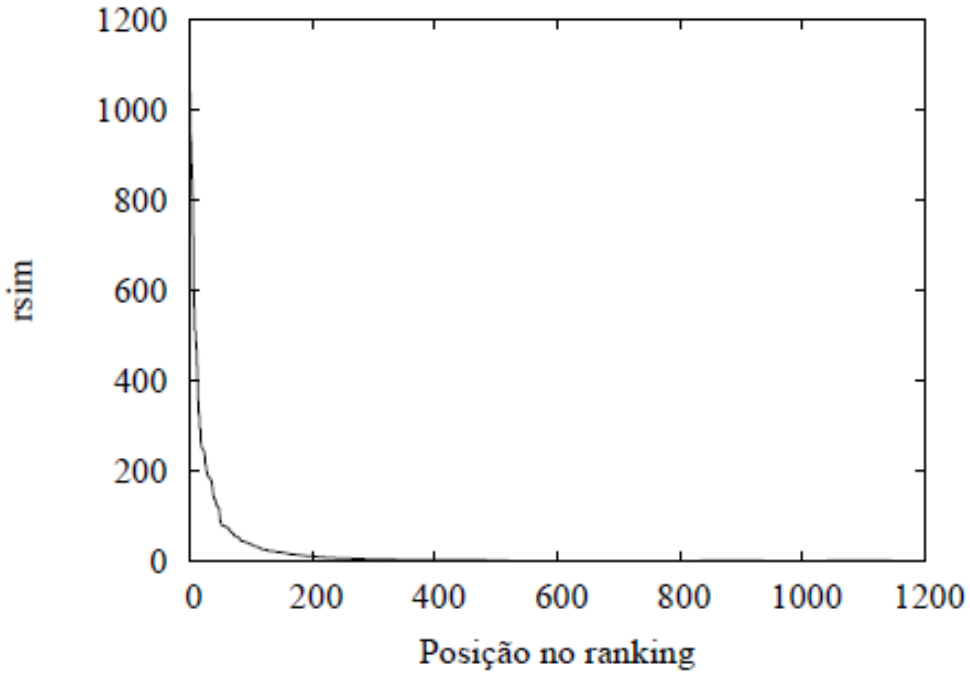
- Pares retornados pela 1ª Fase: **8.906.742**
- O custo do processo de validação ainda é alto
- Tamanho escolhido para o gabarito: **10.000**
  - Amostragem aleatória dos 8.906.742
  - Vale lembrar que a redução no tamanho do gabarito não interfere em nossos objetivos
- Pares de réplicas encontradas pelo validador automático: **583**
- Treinamento e teste dos algoritmos feito sobre o gabarito de 10.000 pares

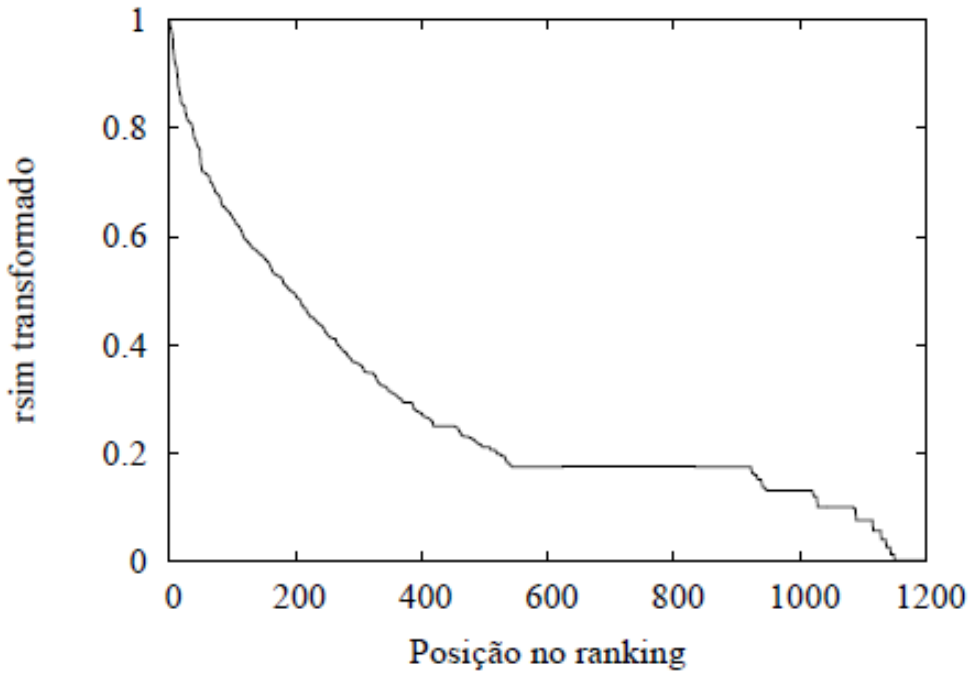
Classes de sítios

- Conteúdo volátil
  - Sítios de ofertas
    - www.ksesporte.com.br e www.materialdenatacao.com.br
- Caminhos trocados
  - Sítios hospedados no mesmo servidor
    - arte.centralblogs.com.br e downloads.centralblogs.com.br
- Conteúdo regional
  - Jornais regionais
    - ac.noticianahora.com.br e ms.noticianahora.com.br
- Avaliação manual de 450 pares de sítios e 15 avaliadores

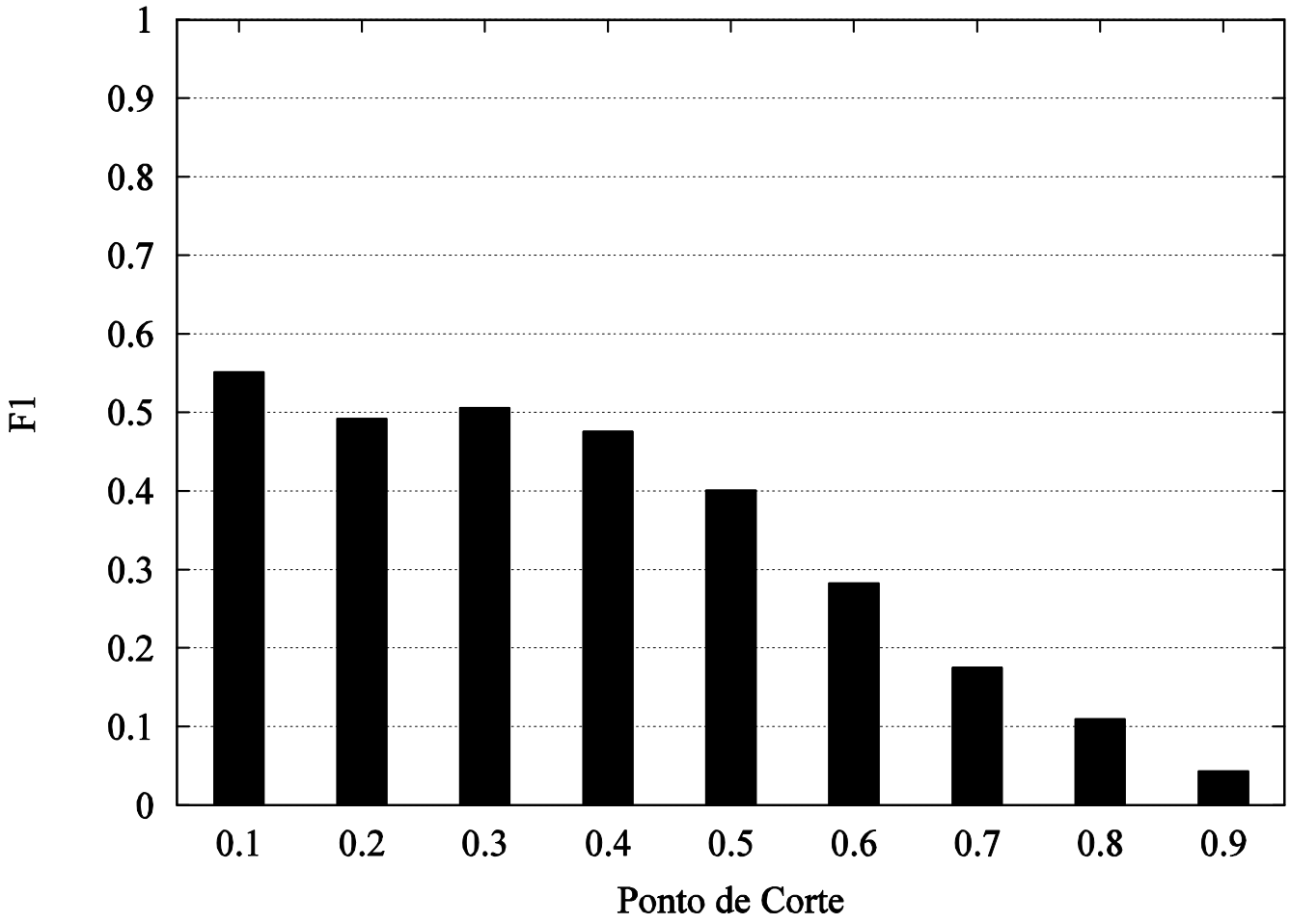
Ajuste de Parâmetros

- Os algoritmos DREAM-*X* retornam uma probabilidade de o par em questão ser réplica
- NormPaths retorna uma pontuação: *rsim*
- *log* e normalização foram aplicados sobre *rsim*

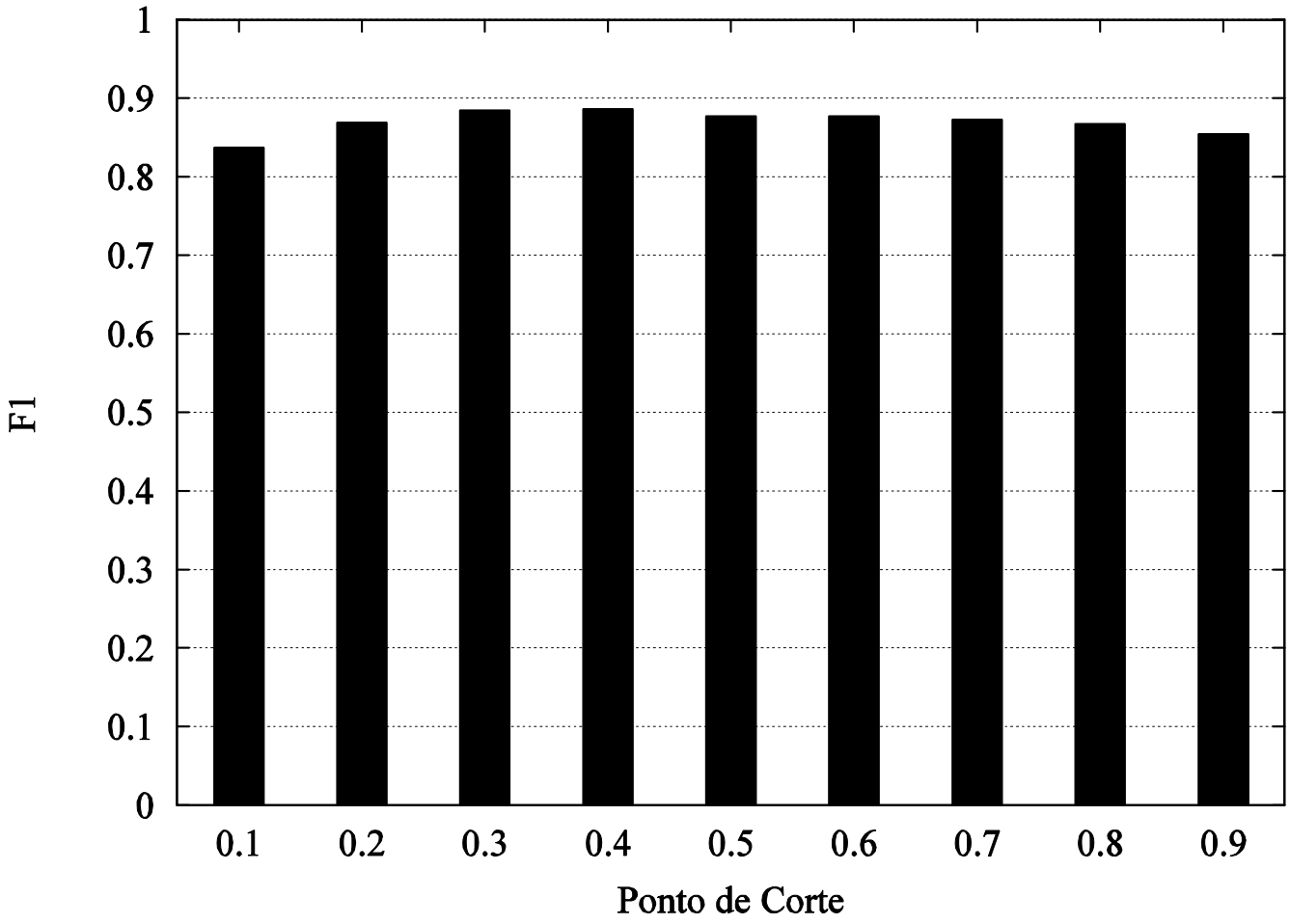




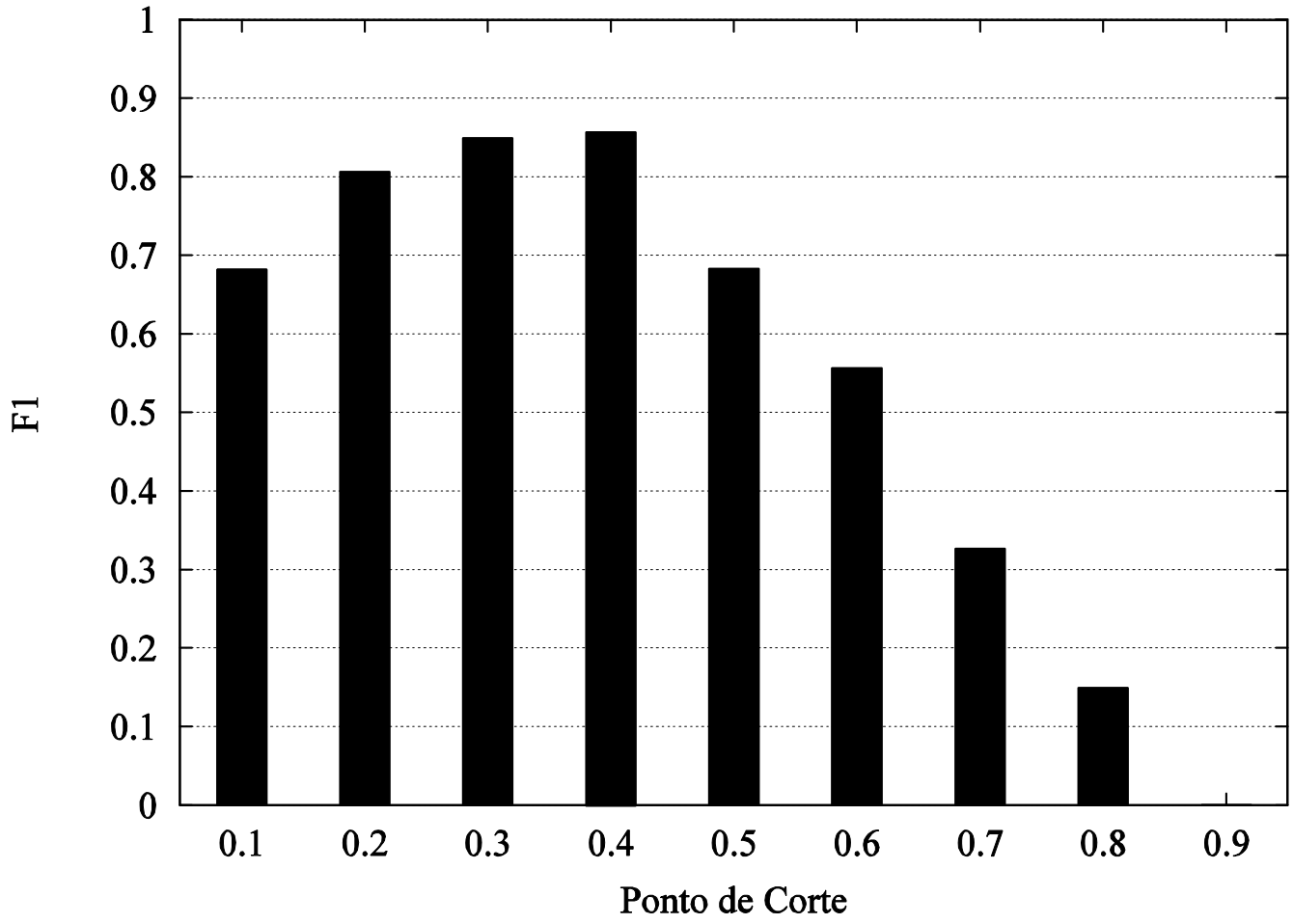
Níveis de F1 do Algoritmo NormPaths



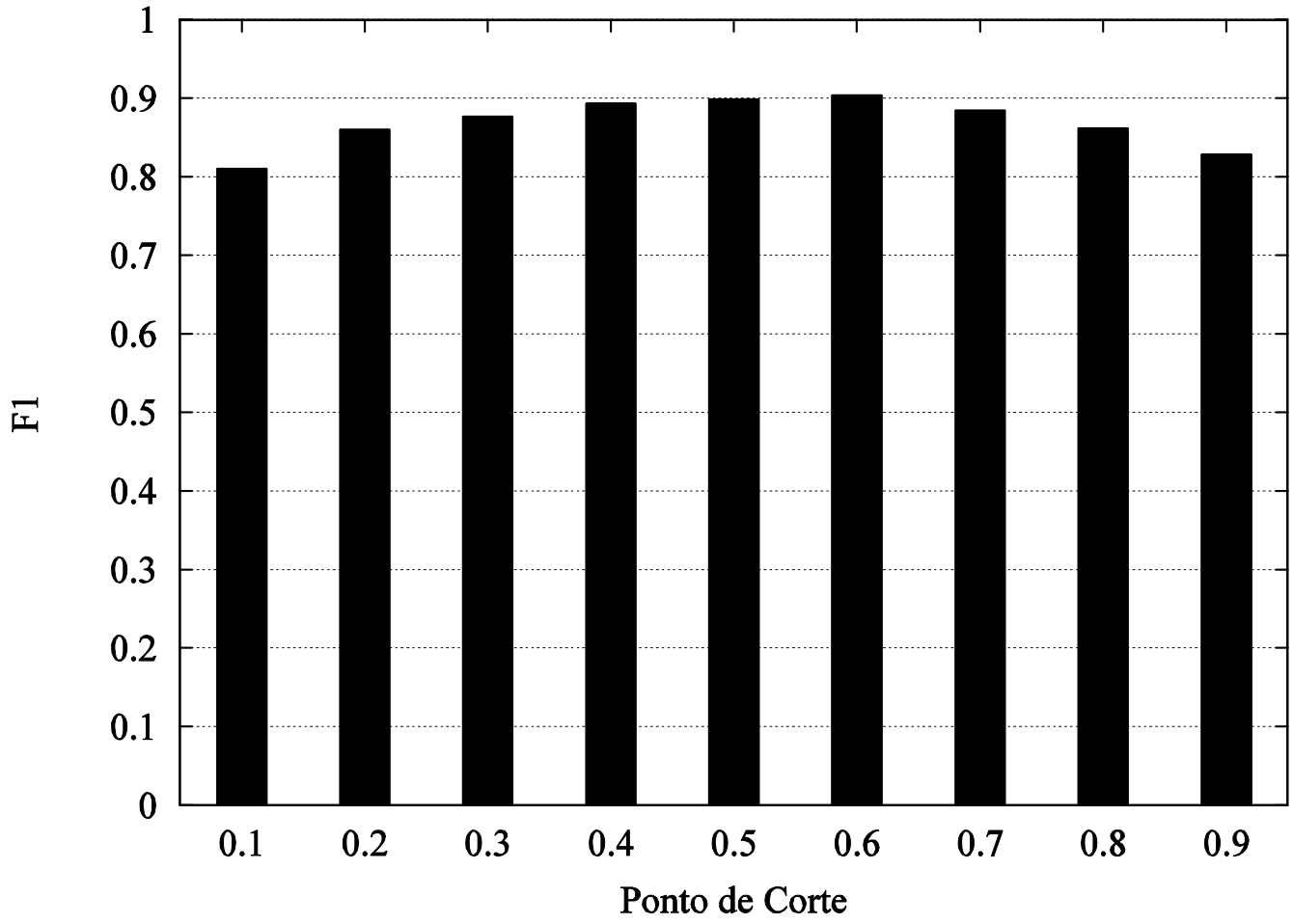
Níveis de F1 do Algoritmo DREAM-DT



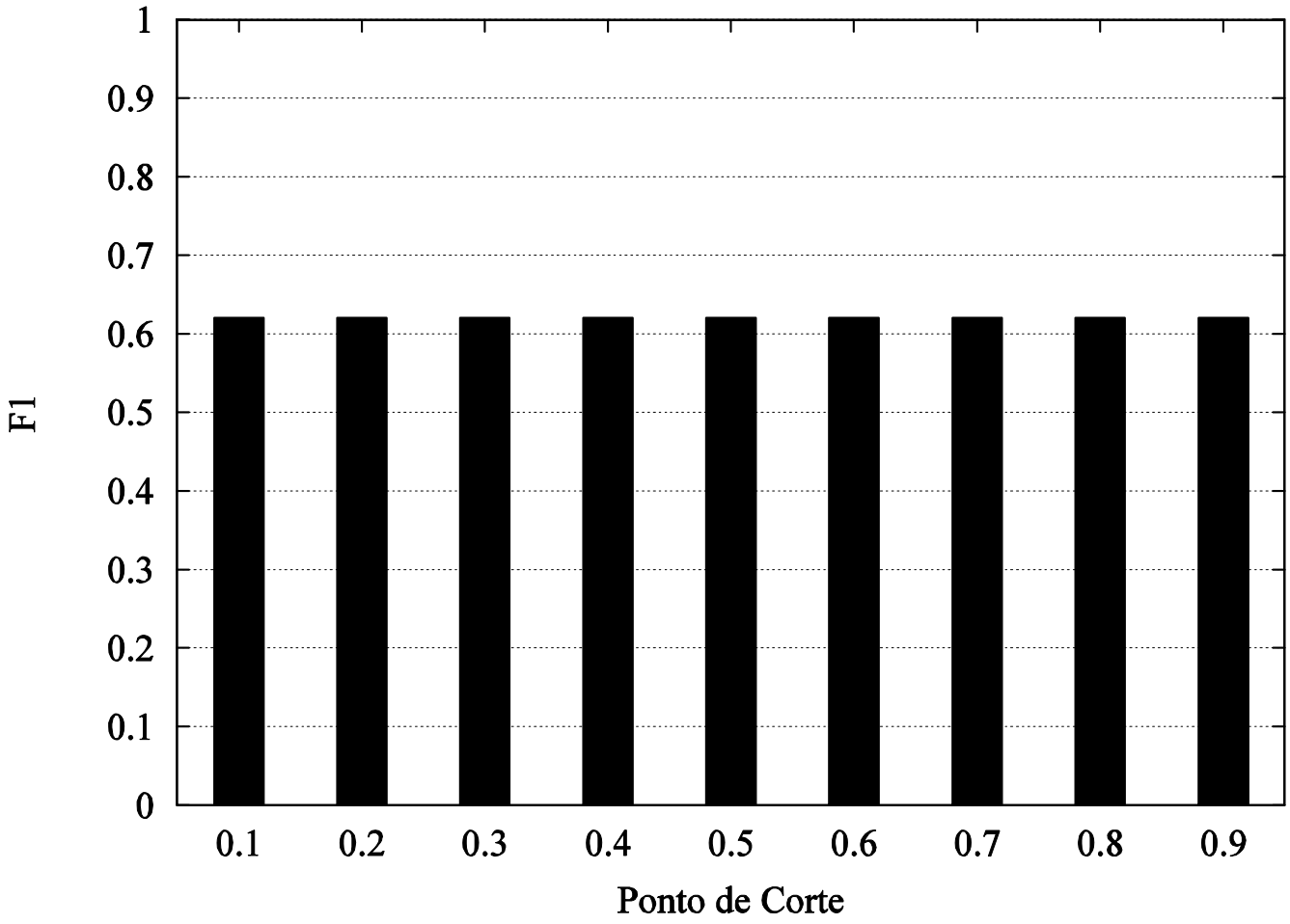
Níveis de F1 do Algoritmo DREAM-LAC



Níveis de F1 do Algoritmo DREAM-RF

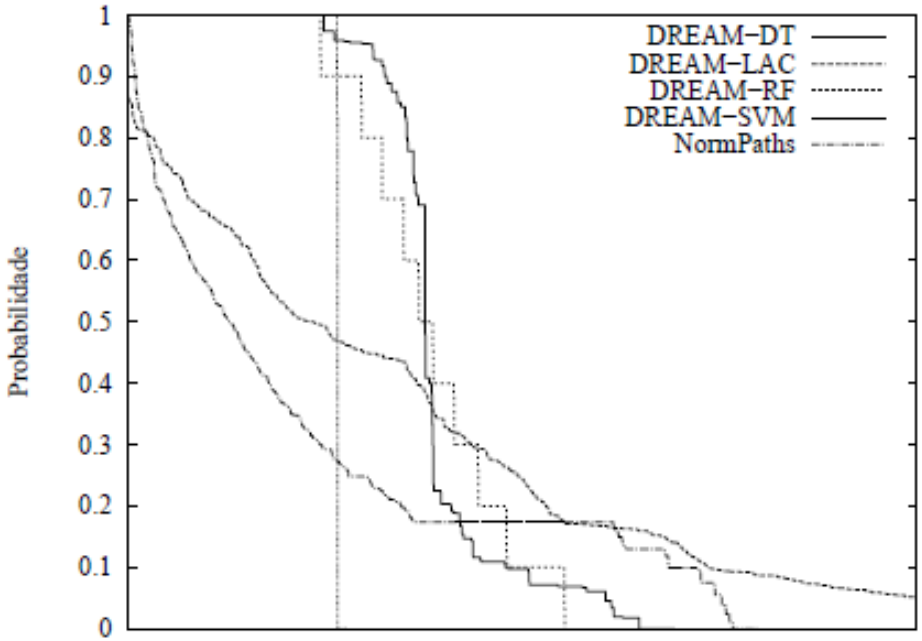


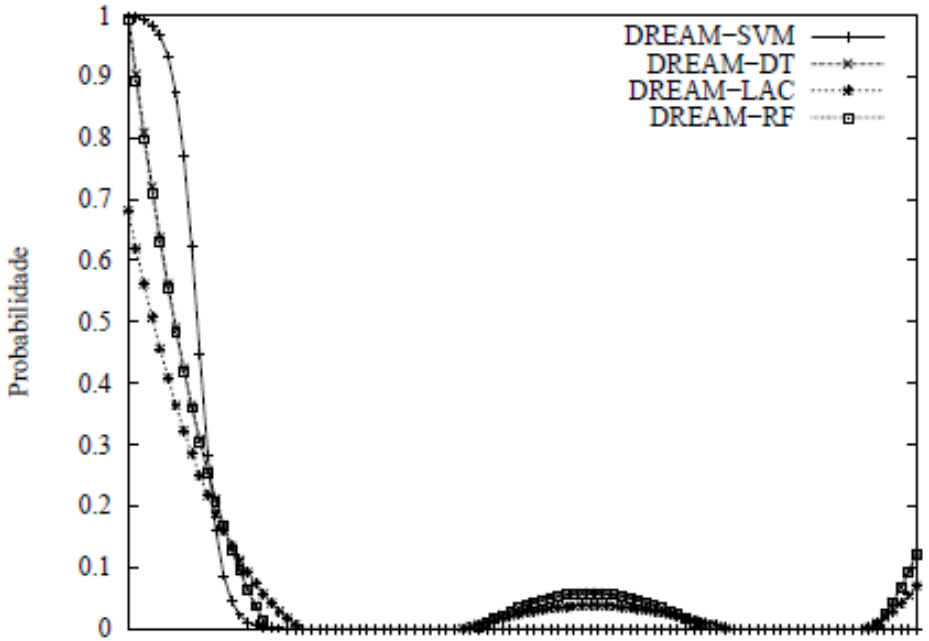
Níveis de F1 do Algoritmo DREAM-SVM



Distribuições de Probabilidade

- DREAM-DT, DREAM-LAC e DREAM-RF
  - Polinômio de quinto grau
- DREAM-SVM
  - Sigmóide: $\frac{1}{1+b\times exp(-a\times x)}$





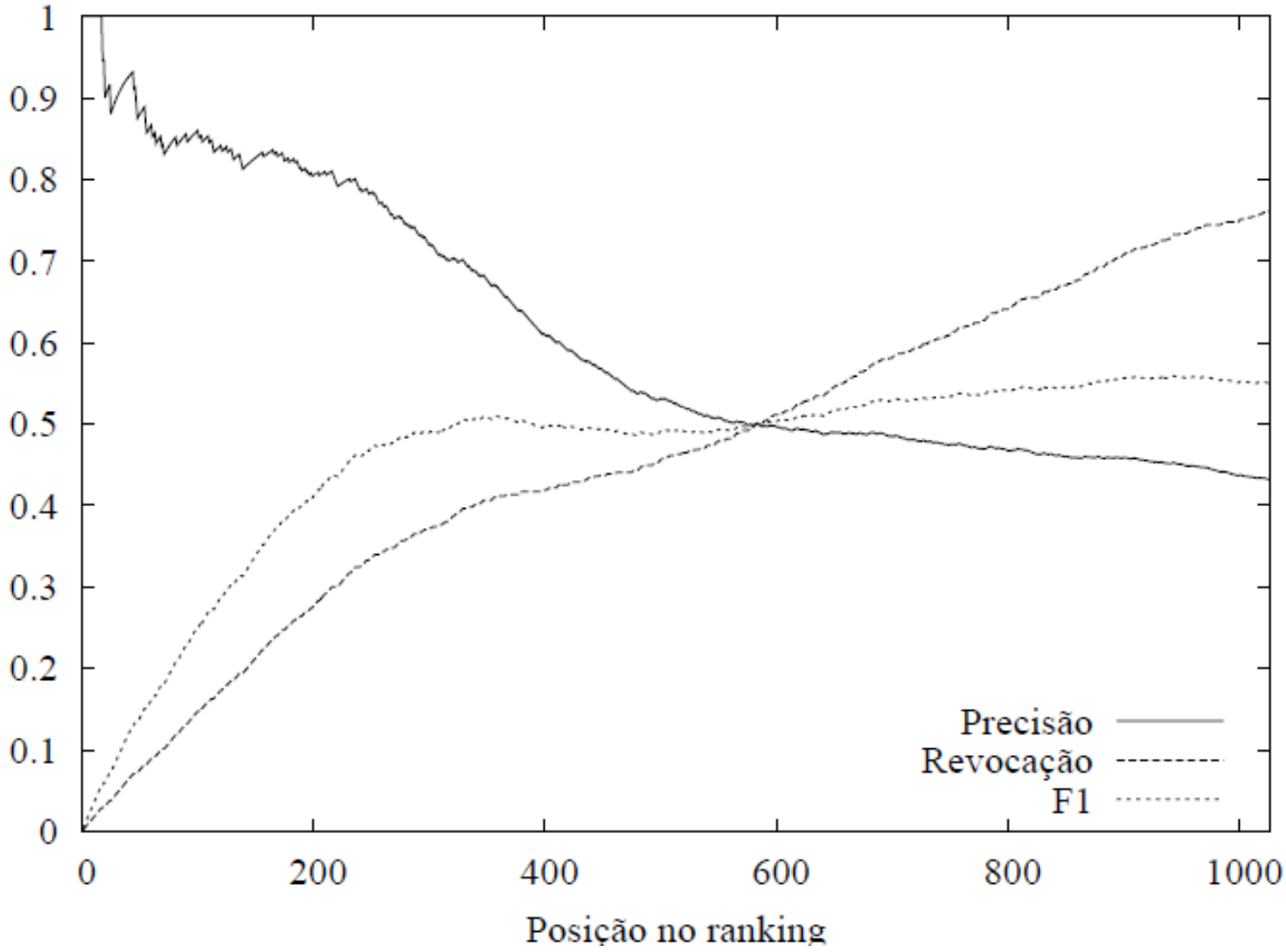
Resultados

- Ajuste de parâmetros

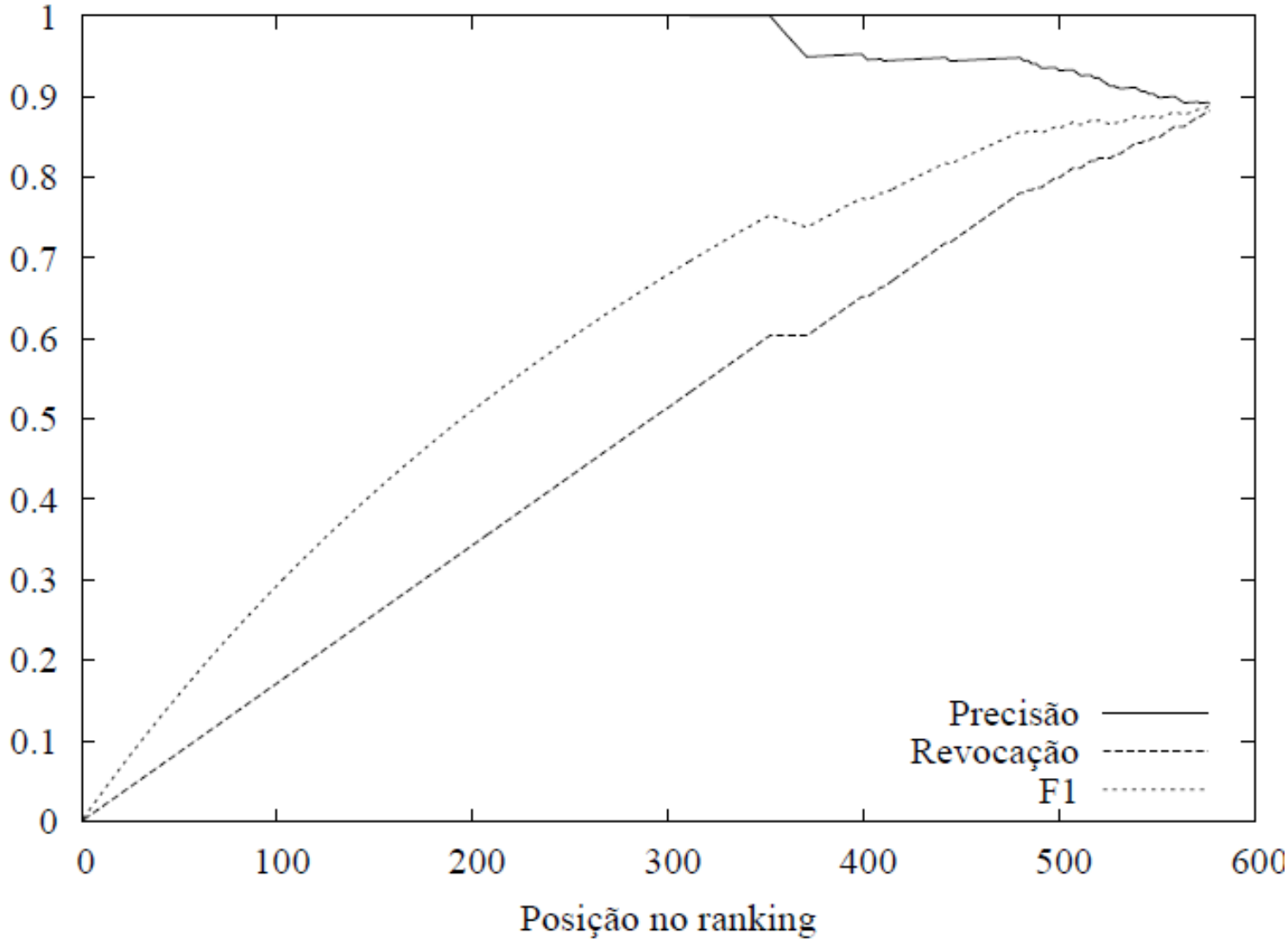
Algoritmo	Ponto de Corte
NormPaths	0,1
DREAM-DT	0,4
DREAM-LAC	0,4
DREAM-RF	0,6
DREAM-SVM	-

- O DREAM-SVM não teve um cenário favorável
  - Todas as probabilidades foram 0 (zero) ou 1 (um)
  - Probabilidades concentradas nos extremos da sigmóide
  - Número insuficiente de características

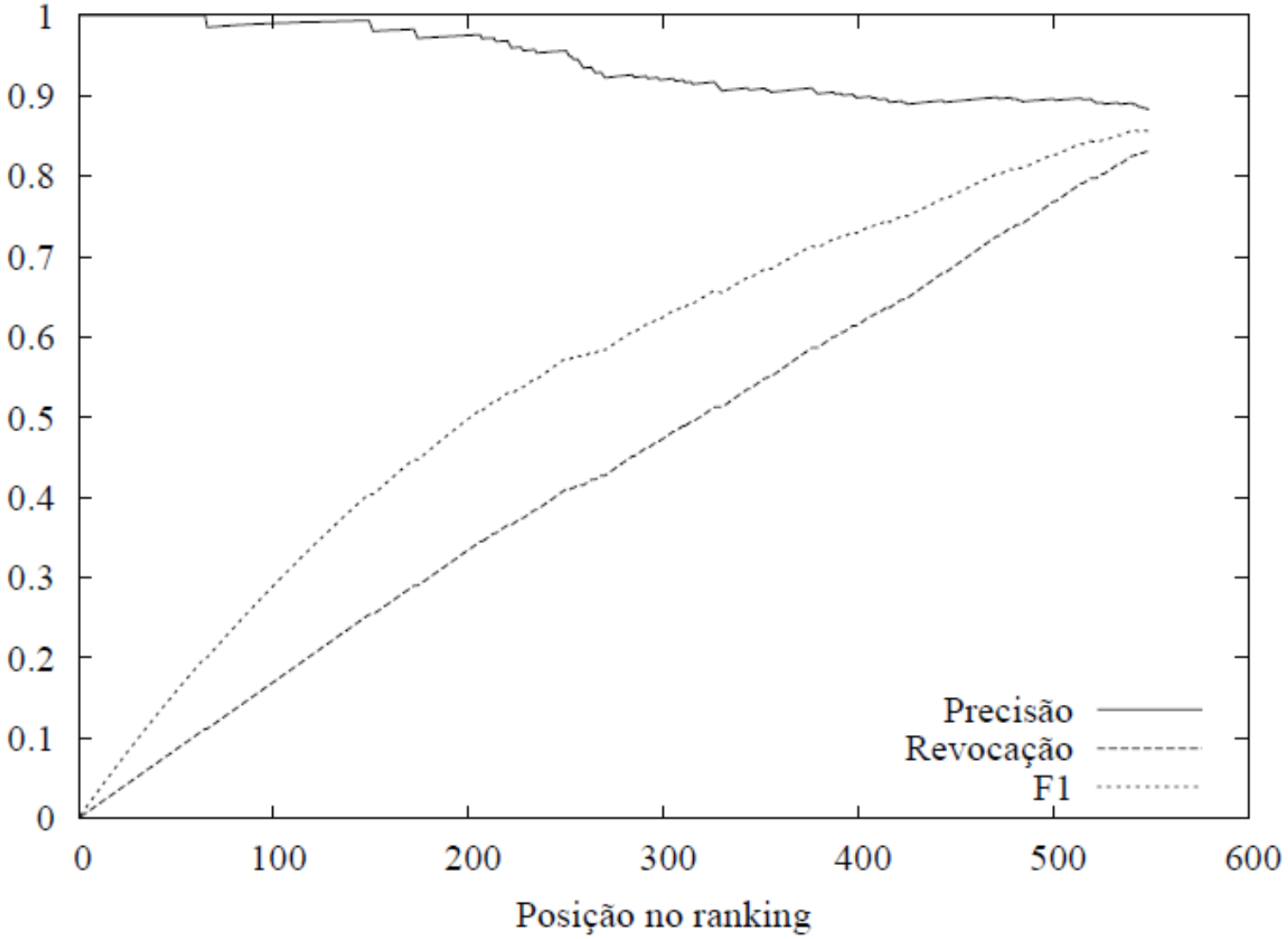
Desempenho do Algoritmo NormPaths



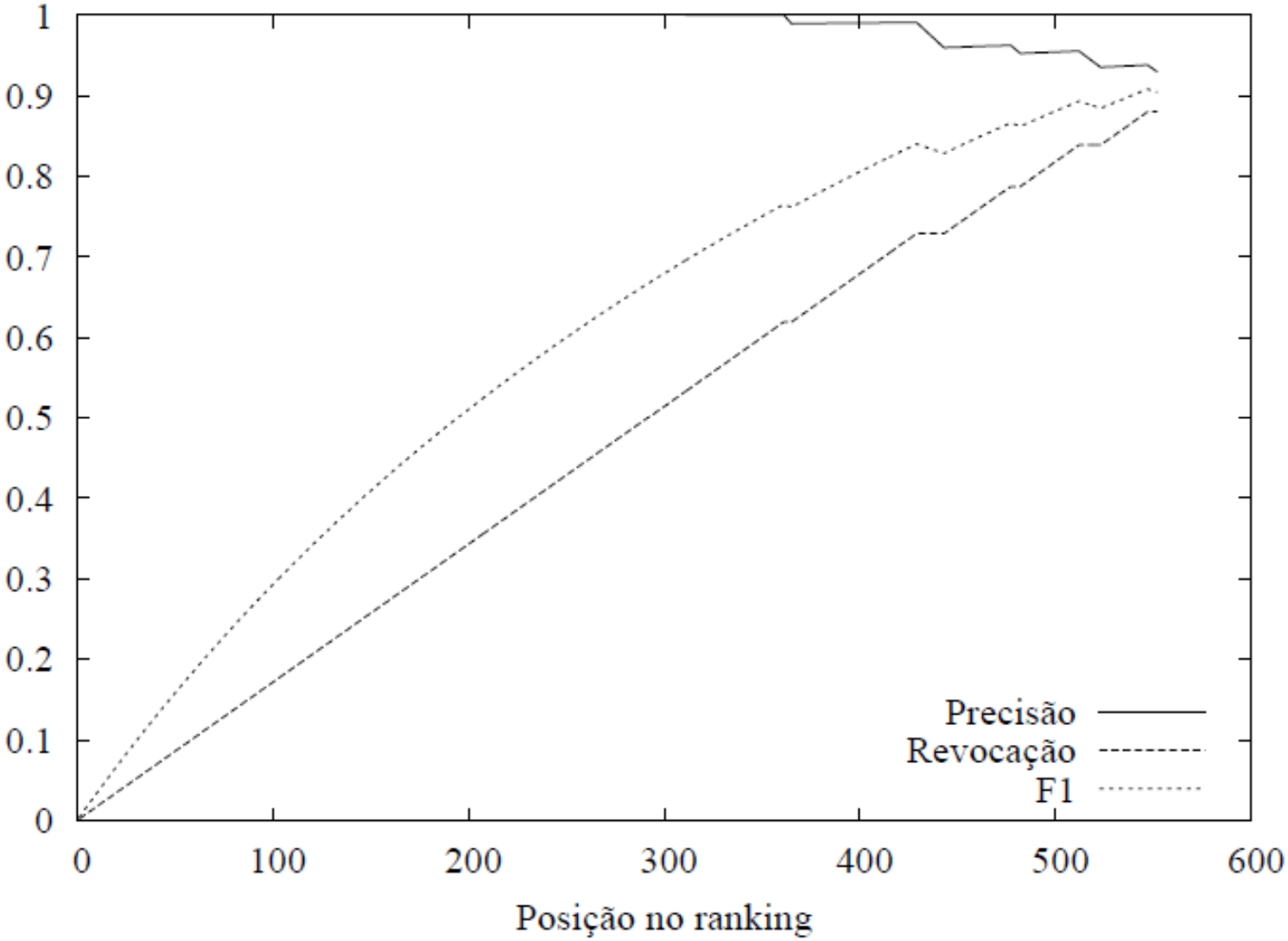
Desempenho do Algoritmo DREAM-DT



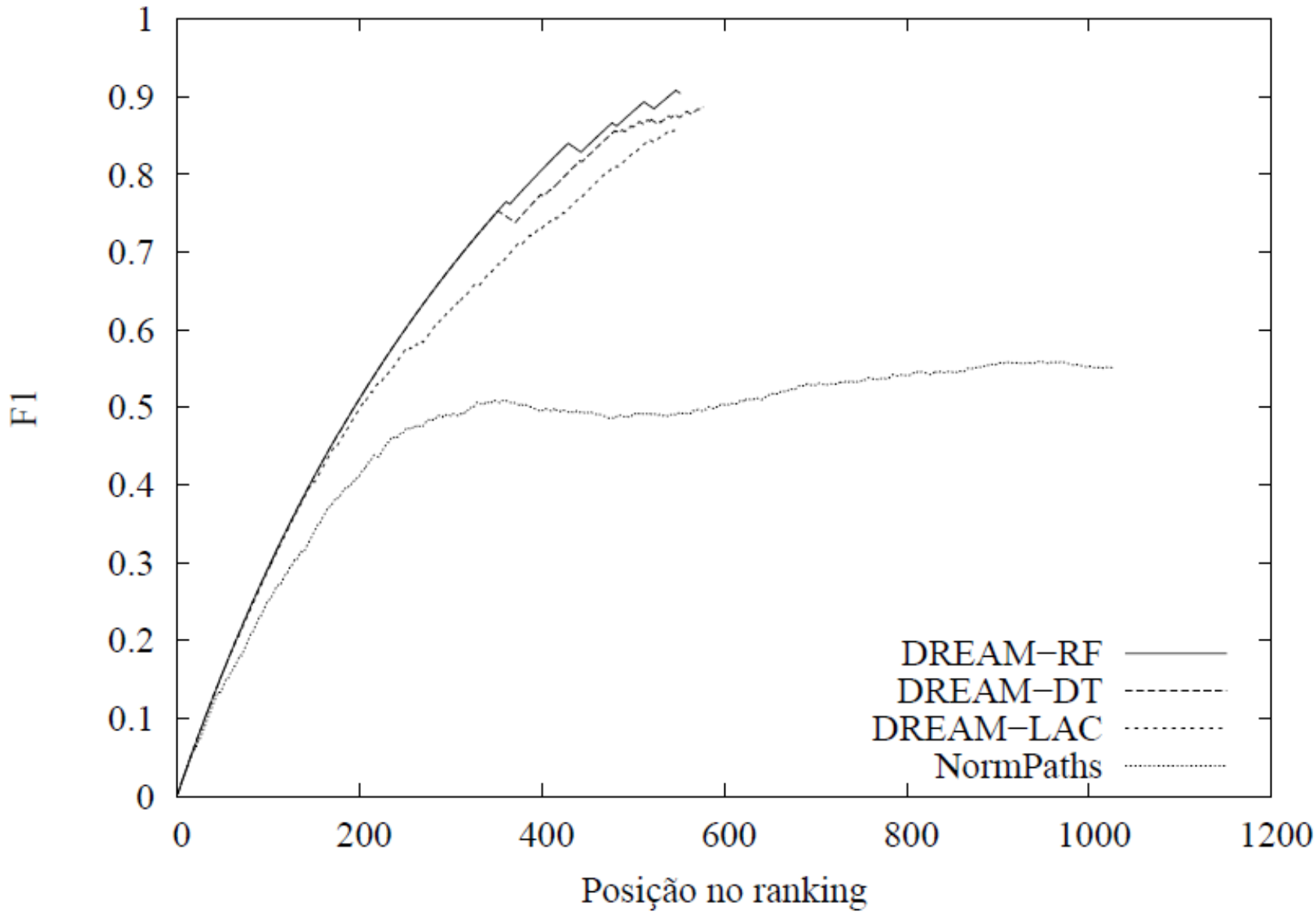
Desempenho do Algoritmo DREAM-LAC



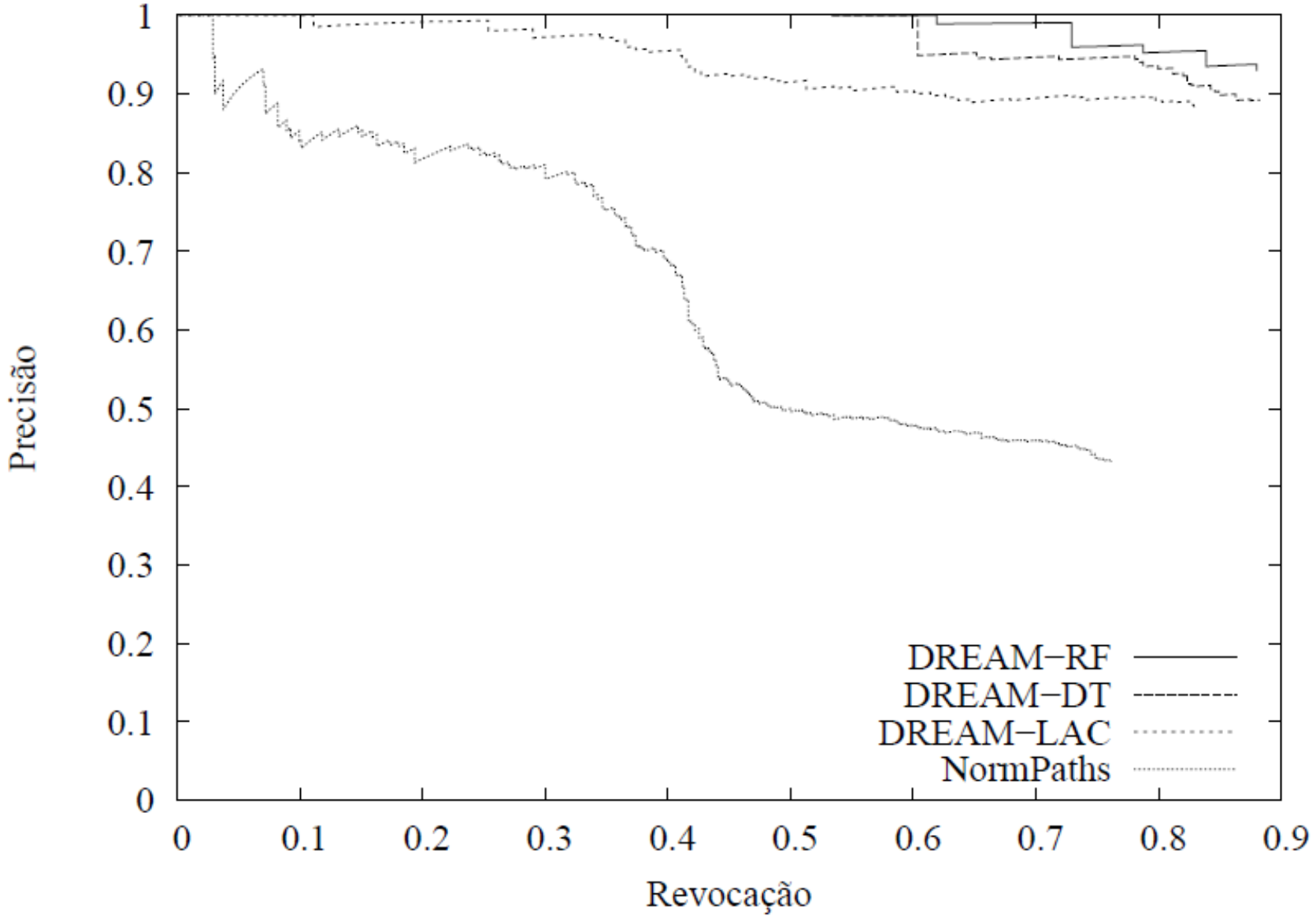
Desempenho do Algoritmo DREAM-RF



Comparação: Curvas de F1



Comparação: Precisão *x* Revocação



Resultados Finais

Algoritmo	Melhor Ponto de Corte	Precisão	Revocação	F1	Ganho
NormPaths	0,1	43,2%	76,2%	55,2%	
DREAM-RF	0,6	92,9%	87,9%	90,3%	35,1%
DREAM-DT	0,4	89,1%	88,2%	88,6%	33,4%
DREAM-LAC	0,4	88,3%	83,2%	85,7%	30,5%
DREAM-SVM	-	76,6%	52,1%	62,0%	6,8%

- Significância estatística dos resultados:
  - DREAM-RF, DREAM-DT e DREAM-LAC empataram
  - DREAM-SVM e NormPaths empataram

Validação Estatística dos Resultados

- 95% de confiança com 5% de erro
- 370 amostras avaliadas manualmente

Método	Precisão
Validador automático	100,0%
DREAM-RF	95,6%
DREAM-DT	91,7%
DREAM-LAC	90,5%
DREAM-SVM	80,0%
NormPaths	45,8%

- Resultados superiores aos obtidos no gabarito

Conclusões

- Apenas o DREAM-SVM não superou o *baseline*
  - Cenário não favorável
- O DREAM conseguiu ganhos acima de 30%
- O DREAM foi capaz de superar os problemas que atrapalharam o *baseline*
  - Uso de várias características
  - Coleção rotulada de pares de réplicas
- Aprendizado de máquina mostrou-se uma abordagem vantajosa ao uso de heurísticas fixas

Trabalhos Futuros

- Aumentar a coleção de dados
- Estudar outras características (conectividade)
- Estudar o impacto individual das características
- Estudo mais completo sobre as classes de sítios
- Avaliar mais métodos de aprendizado de máquina

Contribuições

- Novo algoritmo para detecção de réplicas de sítios web em bases de máquinas de busca
- Estudo de técnicas de aprendizado de máquina como alternativa ao problema de detecção
- Novas características
- Coleção com dados a respeito de replicação
- Estudo sobre classes de sítios capazes de afetar a detecção de réplicas de sítios
- Artigo:
  - A Machine Learning Based Method For Detecting Web Site Replicas

Dúvidas?