

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

Rickson Guidolini

Orientador: Nivio Ziviani

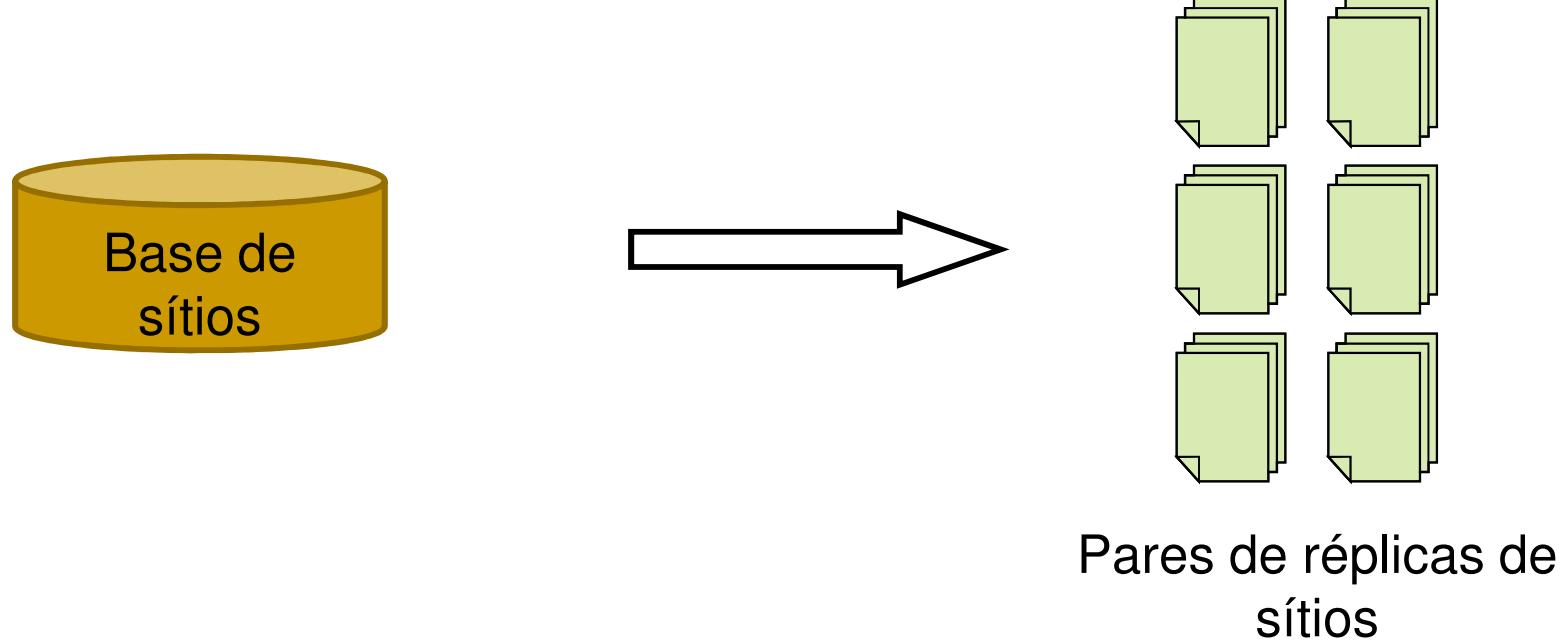
Co-orientador: Adriano Veloso

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

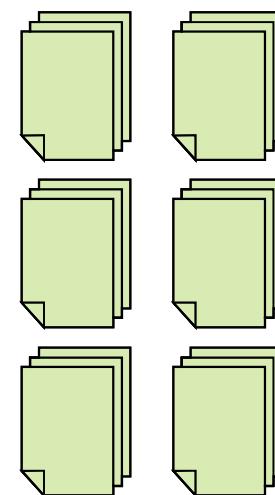
O Problema Abordado



Algoritmo Ingênuo

2.002.955
sítios

$$\frac{n^2 + n}{2}$$



2.005.913.364.535
pares

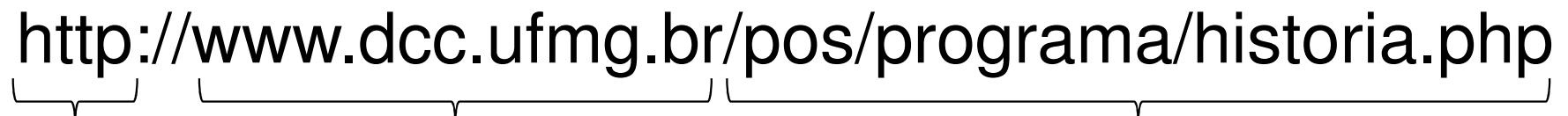
Definição do Problema

- 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

`http://www.dcc.ufmg.br/pos/programa/historia.php`



Método de Acesso Nome de Servidor / sítio Caminho

Custos Envolvidos

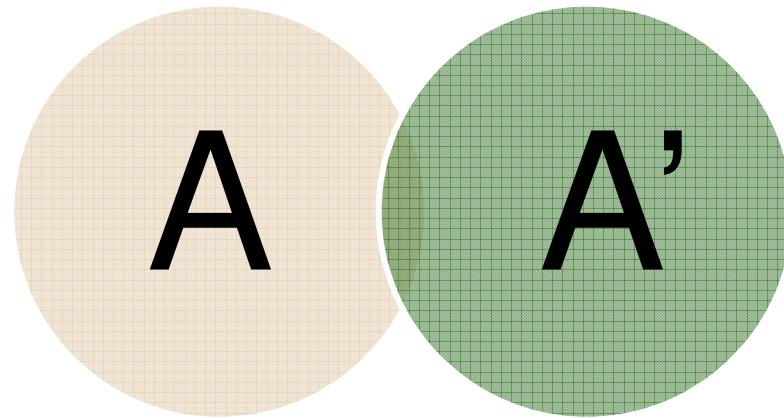
- Verificação de replicação implica em
 - Coleta de páginas web
 - Comparação entre conteúdos textuais
- Para uma amostragem de 10 caminhos:

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

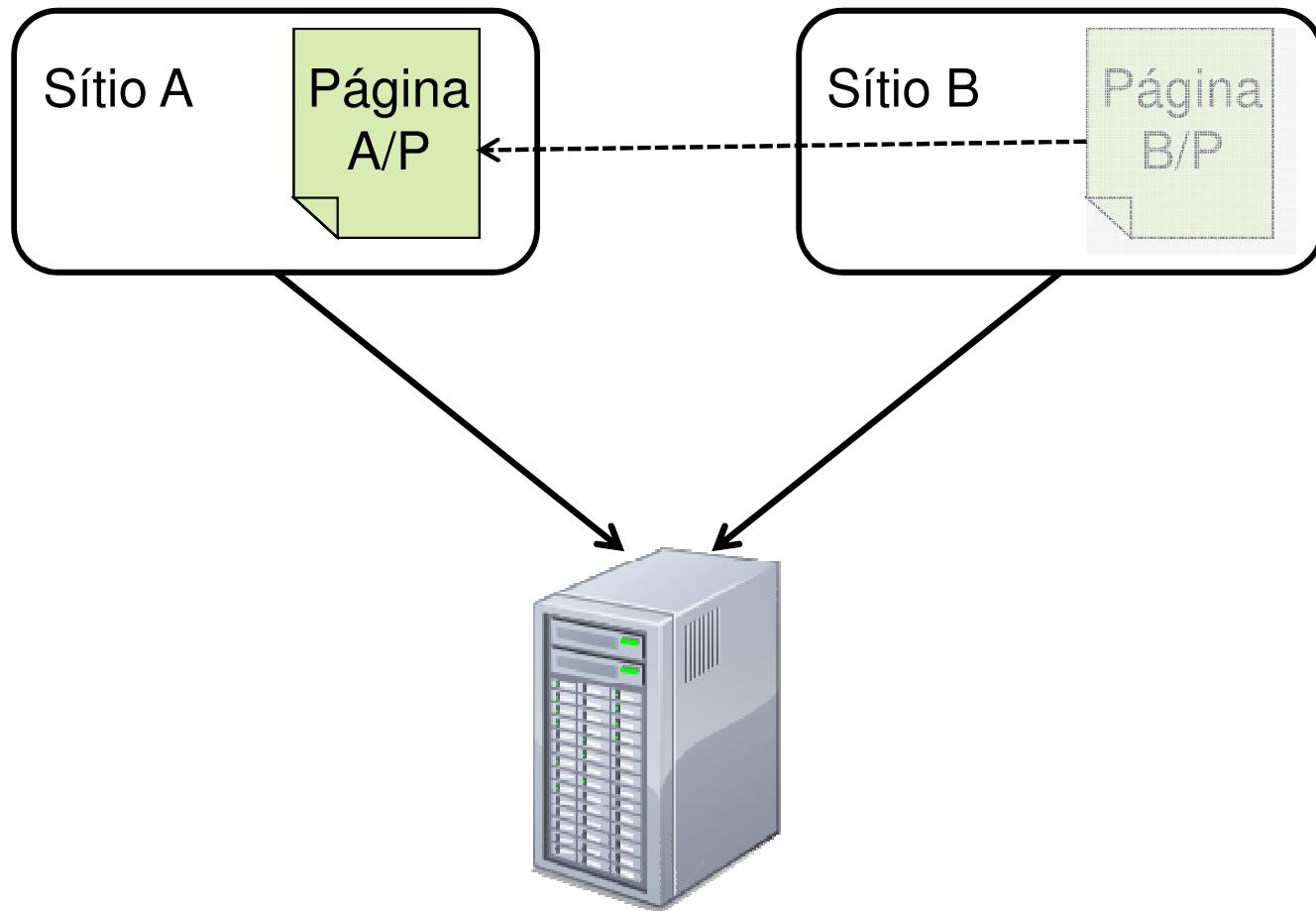
$$comparações = \frac{n^2 + n}{2} \times |amostra| = 20.059.133.645.350$$

Desafios: Cobertura por Sítios

Base de dados da
Máquina de Busca



Desafios: Caminhos Trocados



Desafios: Conteúdo Volátil

WAZ
FANÁTICOS POR TECNOLOGIA

Minha Conta Central de Atendimento wazX Contato Sobre a WAZ

Pesquisa Produto

Fabricantes

Produtos

- Apple
- Bancos e Acessórios
- Cabo/Adaptador
- Caixa de Som
- Câmera Digital
- Cartão de Memória
- Cartucho Impressão
- Casemod
- Computador
- Controladoras
- Cooler
- Cooler CPU
- Cooler Gabinete
- Cooler VGA
- Dissipador
- Diversos
- Drives
- Energia
- Fone de Ouvido
- Fonte
- Gabinete
- Games DS
- Games PC
- Games P53
- Games PSP
- Games Wii
- Games Xbox 360
- Gavetas
- GPS
- Gravador
- HD
- HD Externo
- HD Notebook
- HD SCSI / SAS
- Impressora
- Instrumentos
- Joystick
- KVM
- Memória
- Memória DDR
- Memória DDR2
- Memória DDR3

WAZ Hardware Store

PROMOÇÕES

Bola de Exercício (Powerball) - Newton Digital Led - Laranja/Vermelho - Box

Bola de exercício (Powerball), giroscópio com velocidade de até 15.000rpm, iluminação através de LEDs, display LCD multifunção integrado, aumenta a resistência e fortalece a musculatura das mãos.

R\$ 189,00 à vista
3x R\$ 63,00 sem juros
6x R\$ 33,73

Garantia: 12 meses

Monitor LCD 15,6pol - Samsung B1630N (Widescreen) - Preto - LS16PUYKFLZD - Box

Monitor LCD de 15,6pol (Widescreen), dot pitch 0,252mm, 1.360 x 768, brilho de 250cd/m², contraste de 1.000:1 (dinâmico de 50.000:1), 8ms, conexão VGA, suporte para trava Kensington, pode ser montado na parede (VESA 7,5cm).

R\$ 299,00 à vista
3x R\$ 99,67 sem juros
6x R\$ 53,36

Garantia: 12 meses

Video game - Nintendo Wii - Preto - RVL S KRP2 USZ - Box

Video game Nintendo Wii, suporte para até 4 Wii Remotes, leitor de cartões integrado (SD), suporta áudio em Dolby Pro Logic II, compatível com GameCube, interface WiFi integrada, Wii Remote + Nunchuck + Wii MotionPlus + Wii Sports Resort inclusos.

R\$ 899,00 à vista
3x R\$ 299,67 sem juros
6x R\$ 160,44

Garantia: 6 meses

VGA PCI-E AMD HD 6970 2GB/256bits Sapphire Game Edition - 102-C2001-00-AT / BFBC2 Vietnam - Box

Placa de vídeo com GPU ATI Radeon HD 6970 (880MHz), 2GB de memória (256bits / 5,5GHz), interface PCI Express 16x v2.1, conexões DVI-I + DVI-I (Single Link) + HDMI + 2x Mini DisplayPort, suporta HDCP e HDTV, controlador de áudio (8 canais) integrado, suporta as tecnologias ATI CrossFireX e Eyefinity, game Battlefield: Bad Company 2 Vietnam (Cupom) incluso.

R\$ 1.649,00 à vista
3x R\$ 549,67 sem juros
6x R\$ 294,29

Garantia: 12 meses

Video game portátil - Nintendo DSi XL - Azul/Preto - UTL S BKA USZ - Box

Video game Nintendo DSi XL com dois displays LCD de 4,3pol (sendo um sensível ao toque - stylus inclusa), slot para Game Cards, gamepad, duas câmeras, alto falantes e microfone integrados, interface WiFi integrada, leitor para cartão (SD).

R\$ 779,00 à vista
3x R\$ 259,67 sem juros
6x R\$ 139,02

Garantia: 6 meses

Roteador Wireless - D-Link - DI-524 - 150/BZ - Box

Roteador sem fio de banda larga, compatível com redes sem fio nos padrões 802.11b/g, 2,4GHz, switch integrado com 4 portas Ethernet 10/100, segurança com Firewall e criptografia WEP, WPA e WPA2, suporta NAT, DDNS e VPN Passthrough (PPTP, L2TP e IPsec), uma antena externa.

R\$ 89,00 à vista
3x R\$ 29,67 sem juros
6x R\$ 15,88

Garantia: 36 meses

VGA PCI-E AMD HD 6950 2GB/256bits Sapphire -

Video game portátil - Nintendo DSi - Preto - TWL S KA USZ -

Carrinho de Compras

Nenhum item

O que rola no wazX

[Nova Versão] MSI Afterburner v2.1.0

WAZ

Siga-nos...

twitter

Entre em nossa comunidade...

orkut

Visite nosso blog...

WAZ BLOG

Páginas dos Produtos

LITEON

Desafios: Conteúdo Regional



Meu iClaz | Entre ou Cadastre-se Já

Todo o site

Pesquisas mais populares: nokia - dvd carro - wii - all star - sony vaio - mp4 - notebook - xbox - iphone 3g - blazer - santana -

Localização: Brasil > Abelardo Luz | Classificados Grátis em Abelardo Luz

Localização	
Selecionar uma Cidade	
São Paulo	Porto Alegre
Rio de Janeiro	Fortaleza
Belo Horizonte	Campinas
Londrina	Florianópolis
Curitiba	Goiânia
Brasília	Niterói
Santo André	Limeira
Guarulhos	São Caetano do Sul
Salvador	Sul
São Bernardo do Campo	Bauru
São Paulo	São João de Meriti
Selecionar um Estado	
Acre	Pará
Alegoas	Paraná
Amapá	Pernambuco
Amazonas	Piauí
Bahia	Rio de Janeiro
Ceará	Rio Grande do Norte
Distrito Federal	Rio Grande do Sul
Espírito Santo	Rondônia
Maranhão	Roraima
Mato Grosso	Santa Catarina
Mato Grosso do Sul	São Paulo
Minas Gerais	Sergipe
Pará	Tocantins

Loja Virtual

Tenha o Pagamento Digital Exclusivo e Ganhe os Cliques de sua Campanha!

[Negozi.Buzzpage.com.br/LojaVirtual](#)

Seu Site na 1ª Página

Garantimos seu Site na 1ª Página dos Buscadores em Poucos Dias !

[Custamundo.com.br](#)

[Andamento](#)

Categorias (3796 anúncios)

- [Compras - Venda \(2\)](#)
- [Carros, motos e barcos \(1\)](#)
- [Empregos \(0\)](#)
- Animais domésticos - Estimação
- Arte - Coleções - Antigüidades
- Artigos Esportivos - Bicicletas
- Casa - Jardim - Móveis
- Celulares - Acessórios
- Computadores - Informática
- Discos Vinyl - CDs - Música
- DVD - Filmes
- Eletrônicos
- Entradas - Ingressos
- Fotografia - Imagem - Som
- Instrumentos Musicais
- Jogos - Brinquedos
- Jóias - Bijuteria - Relógios
- Livros - Revistas
- Para Bebês - Crianças
- Permuta - Trocas
- Roupa - Acessórios - Moda
- Saúde - Beleza
- Videogames - Consoles
- Outras Compras - Vendas
- [Serviços \(2\)](#)
- [Cursos - Aulas \(0\)](#)
- Cursos de Línguas
- Música - Teatro - Dança
- Reforço - Aulas Particulares
- Web design - Multimídia
- Outros Cursos - Aulas
- Babá
- Horóscopo - Tarô
- Modelos - Casting
- Organização de Eventos
- Reparação
- Saúde - Beleza
- Serviços de Informática
- [Carros, motos e barcos \(1\)](#)
- Carros
- Peças e Acessórios
- Motocicletas - Scooters
- Barcos - Lanchas
- Trailers - Reboques - Caravanas
- Caminhões - Veículos Comerciais
- Carcasas
- Outros Veículos
- [Imóveis \(0\)](#)
- Apartamento - Casa à Venda
- Apartamento - Casa para Aluguel
- Quartos em Aluguel - Companheiros de Quarto
- Casas para Trocar
- Marketing - Publicidade
- Medicina - Enfermagem
- Publicidade - Relações Públicas
- Recursos Humanos
- Tecnologia - Informática - Programação
- Trabalho Voluntário
- Varejo
- Vendas
- Outros Empregos
- [Encontros \(0\)](#)
- Mulher procura Homem
- Homem procura Mulher
- Homem procura Homem
- Mulher procura Mulher



Meu iClaz | Entre ou Cadastre-se Já

Todo o site

Pesquisas mais populares: blazer - dvd carro - parati - jeep - relógio - bike - fusca - pionner - playstation - golf - violao -

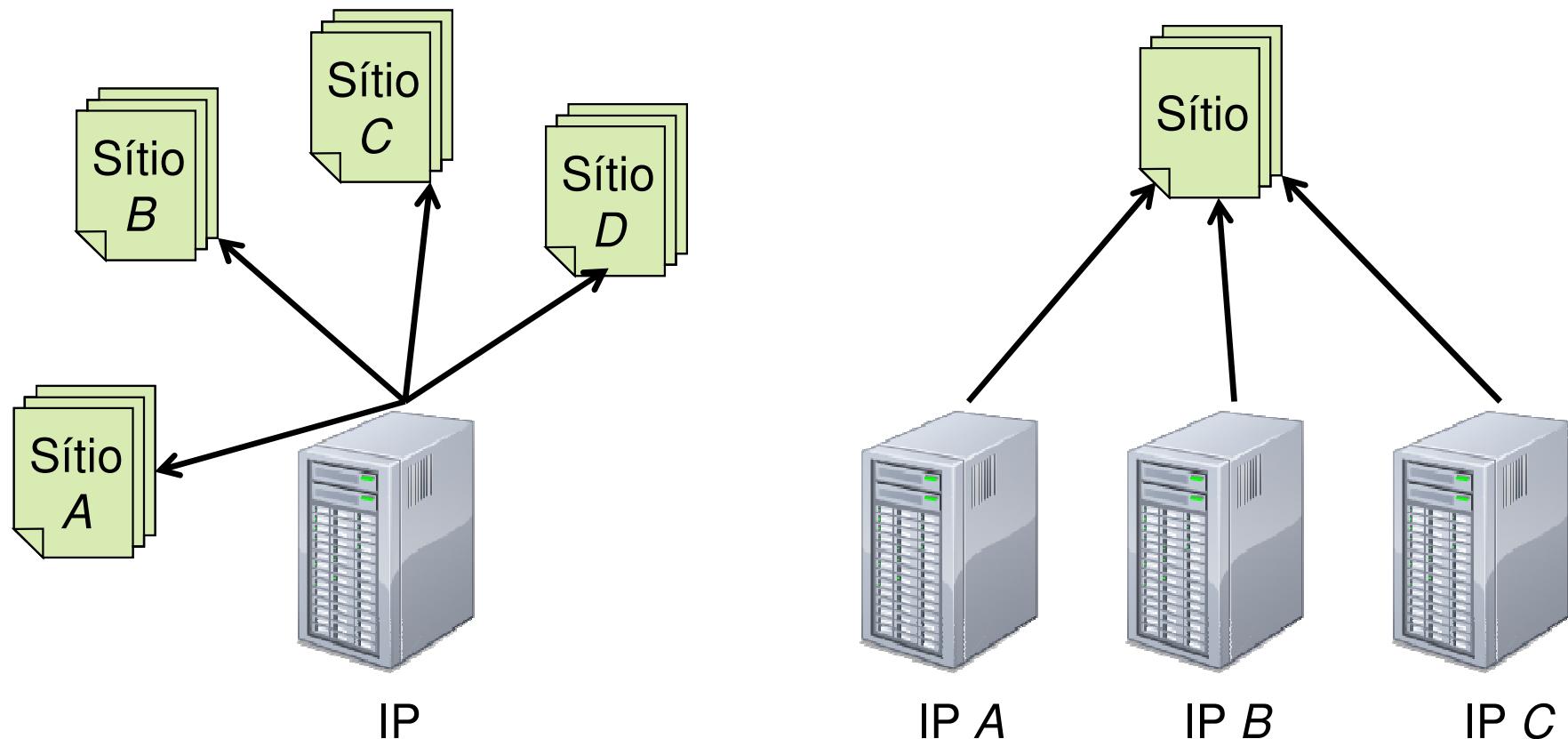
Localização: Brasil > São Paulo | Classificados Grátis em São Paulo

Categorias (7775 anúncios)

- [Compras - Venda \(1572\)](#)
- [Carros, motos e barcos \(312\)](#)
- [Empregos \(272\)](#)
- Animais domésticos - Estimação
- Arte - Coleções - Antigüidades
- Artigos Esportivos - Bicicletas
- Casa - Jardim - Móveis
- Celulares - Acessórios
- Computadores - Informática
- Discos Vinyl - CDs - Música
- DVD - Filmes
- Eletrônicos
- Entradas - Ingressos
- Fotografia - Imagem - Som
- Instrumentos Musicais
- Jogos - Brinquedos
- Jóias - Bijuteria - Relógios
- Livros - Revistas
- Para Bebês - Crianças
- Permuta - Trocas
- Roupa - Acessórios - Moda
- Saúde - Beleza
- Videogames - Consoles
- Outras Compras - Vendas
- [Serviços \(434\)](#)
- [Imóveis \(343\)](#)
- Apartamento - Casa à Venda
- Apartamento - Casa para Aluguel
- Quartos em Aluguel - Companheiros de Quarto
- Casas para Trocar
- Aluguel por Temporada
- Vagas de Estacionamento
- Terrenos
- Escritórios - Locais de Comércio
- Pontos Comerciais para Alugar - Vender
- [Carros, motos e barcos \(312\)](#)
- Carros
- Peças e Acessórios
- Motocicletas - Scooters
- Barcos - Lanchas
- Trailers - Reboques - Caravanas
- Caminhões - Veículos Comerciais
- Contabilidade - Finanças
- Direito - Advocacia
- Educação - Professores
- Engenharia - Arquitetura
- Hotéis - Turismo - Restaurante
- Imobiliária
- Indústria
- Internet - Multimídia
- Marketing - Publicidade
- Medicina - Enfermagem
- Publicidade - Relações Públicas
- Recursos Humanos
- Tecnologia - Informática - Programação
- Trabalho Voluntário
- Varejo
- Vendas
- Outros Empregos
- [Encontros \(217\)](#)
- Mulher procura Homem
- Homem procura Mulher
- Homem procura Homem
- Mulher procura Mulher
- [Cursos - Aulas \(331\)](#)
- Cursos de Línguas
- Música - Teatro - Dança
- Reforço - Aulas Particulares
- Web design - Multimídia
- Outros Cursos - Aulas
- Babá
- Horóscopo - Tarô
- Modelos - Casting
- Organização de Eventos
- Reparação
- Saúde - Beleza
- Serviços de Informática
- [Empregos \(272\)](#)
- Administrador - Secretária - Setor Público
- Animador para festas e eventos - Ator
- Atendimento ao Cliente
- Construção Civil
- Contabilidade - Finanças
- Direito - Advocacia
- Educação - Professores
- Engenharia - Arquitetura
- Hotéis - Turismo - Restaurante
- Imobiliária
- Indústria
- Internet - Multimídia
- Marketing - Publicidade
- Medicina - Enfermagem
- Publicidade - Relações Públicas
- Recursos Humanos
- Tecnologia - Informática - Programação
- Trabalho Voluntário
- Varejo
- Vendas
- Outros Empregos
- [Loca Virtual](#)
- Tenha o Pagamento Digital Exclusivo e Ganhe os Cliques de sua Campanha!
- [www.GROUPON.com.br/mais](#)
- [Loca Virtual](#)
- Tenha o Pagamento Digital Exclusivo e Ganhe os Cliques de sua Campanha!
- [www.GROUPON.com.br/mais](#)

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

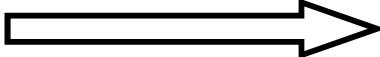
- Fase 1: Características gerais são utilizadas para selecionar um conjunto de pares com algum potencial para serem réplicas
- Fase 2: Características específicas são utilizadas pelo modelo aprendido (aprendizado de máquina), obtendo uma lista refinada de pares de sítios
- Fase 3: Respostas do modelo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

Características Utilizadas

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

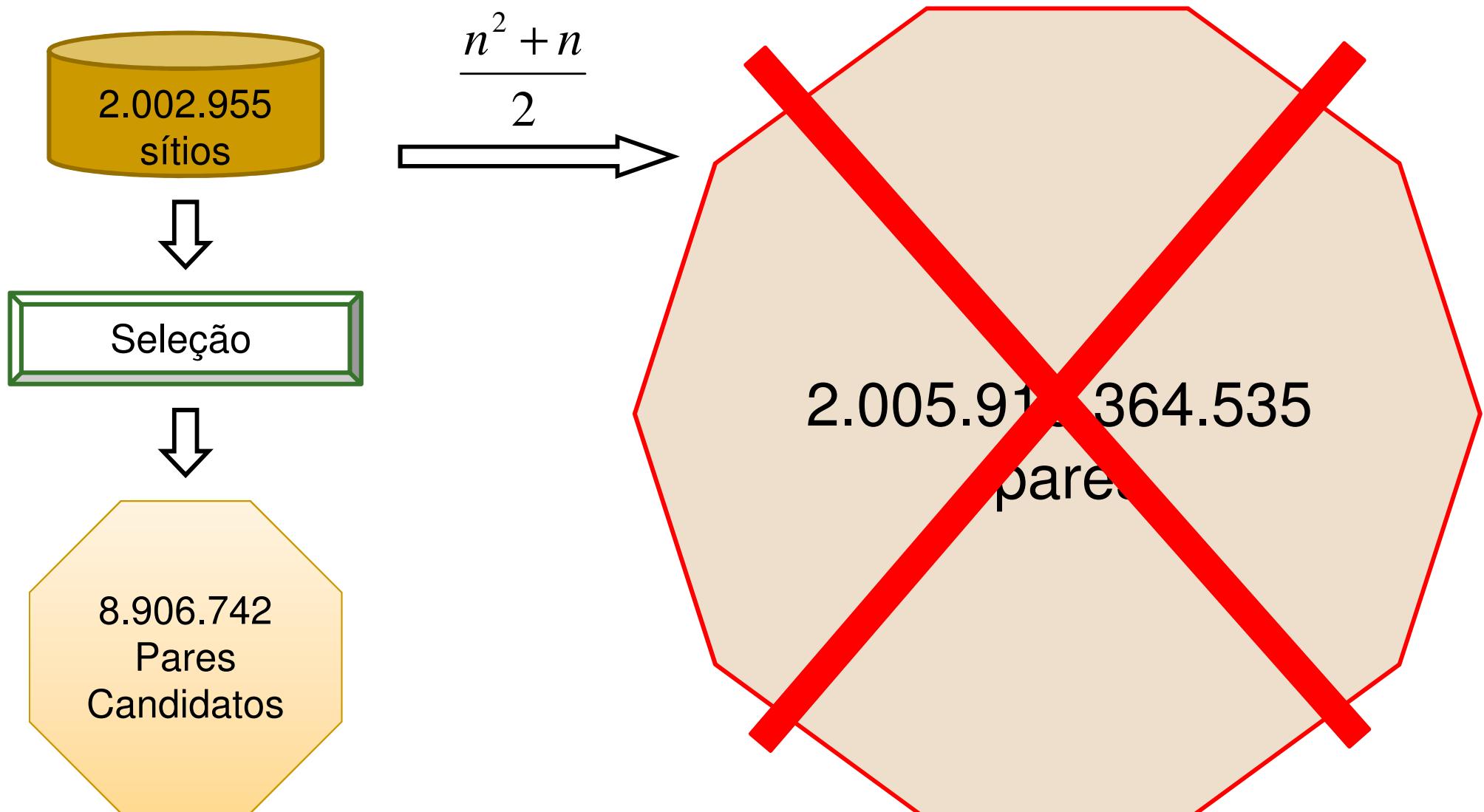
Algoritmo Ingênuo

2.002.955
sítios

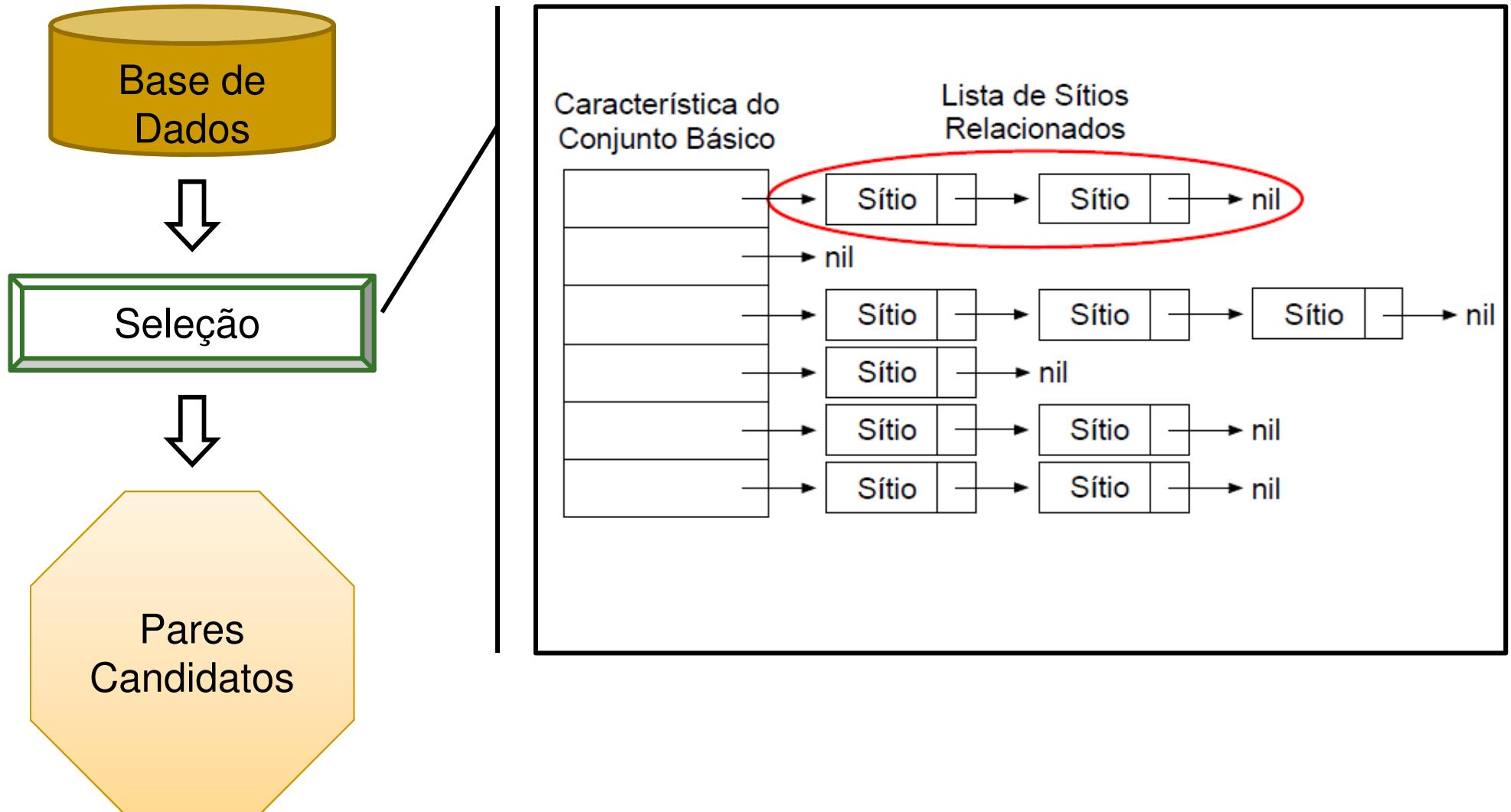
$$\frac{n^2 + n}{2}$$


2.005.913.364.535
pares

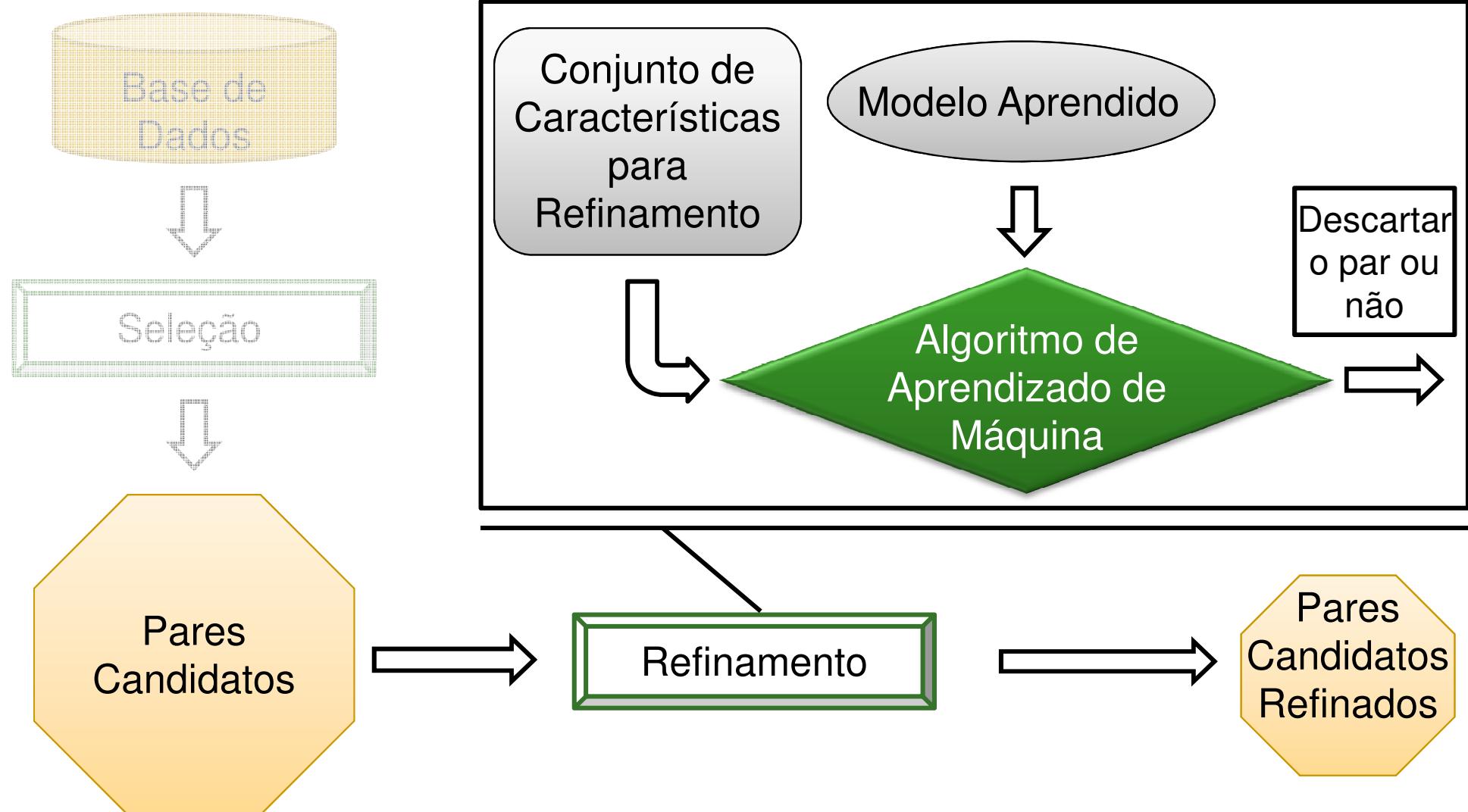
1^a Fase versus Algoritmo Ingênuo



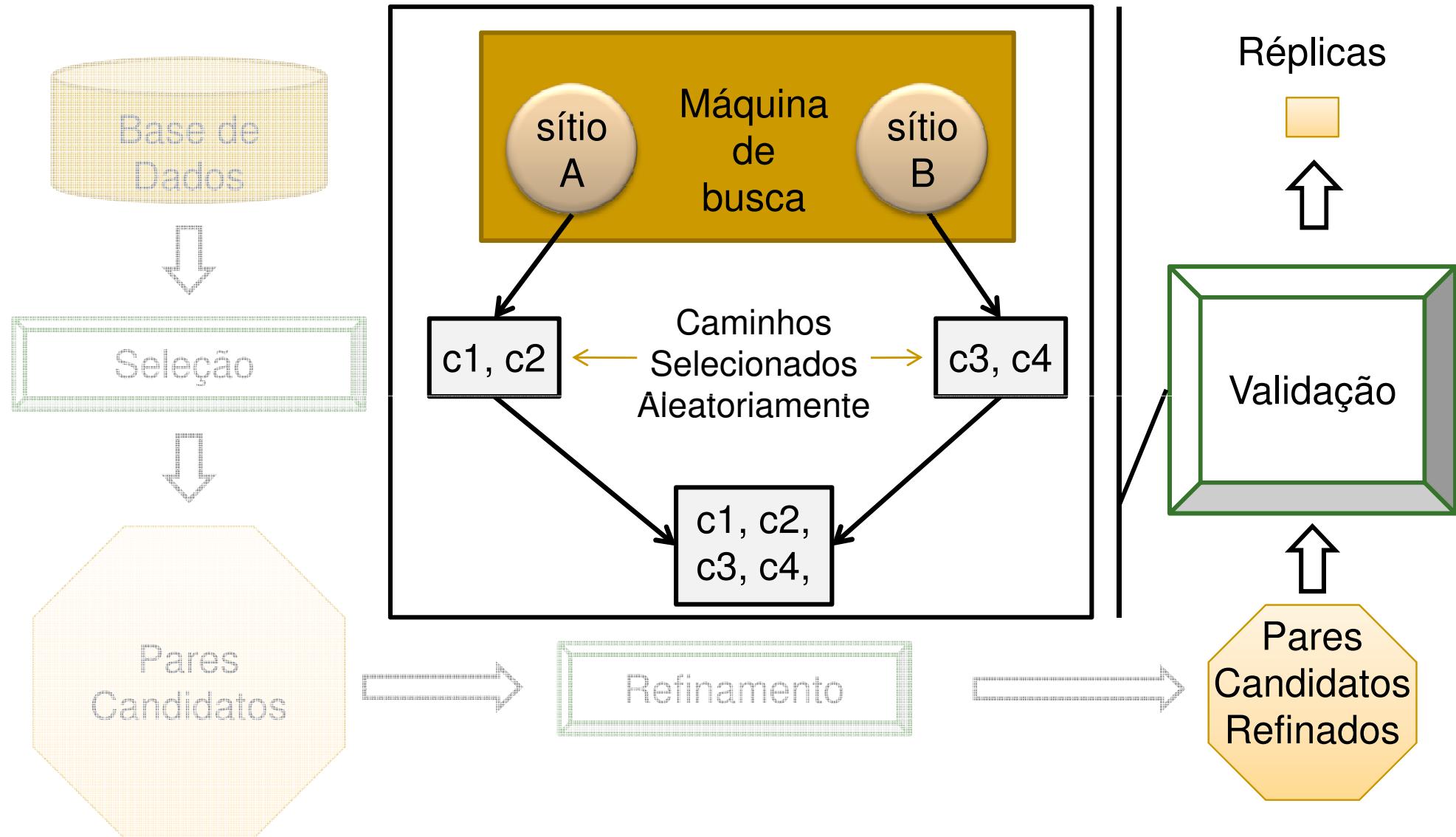
1^a Fase: Seleção de pares candidatos



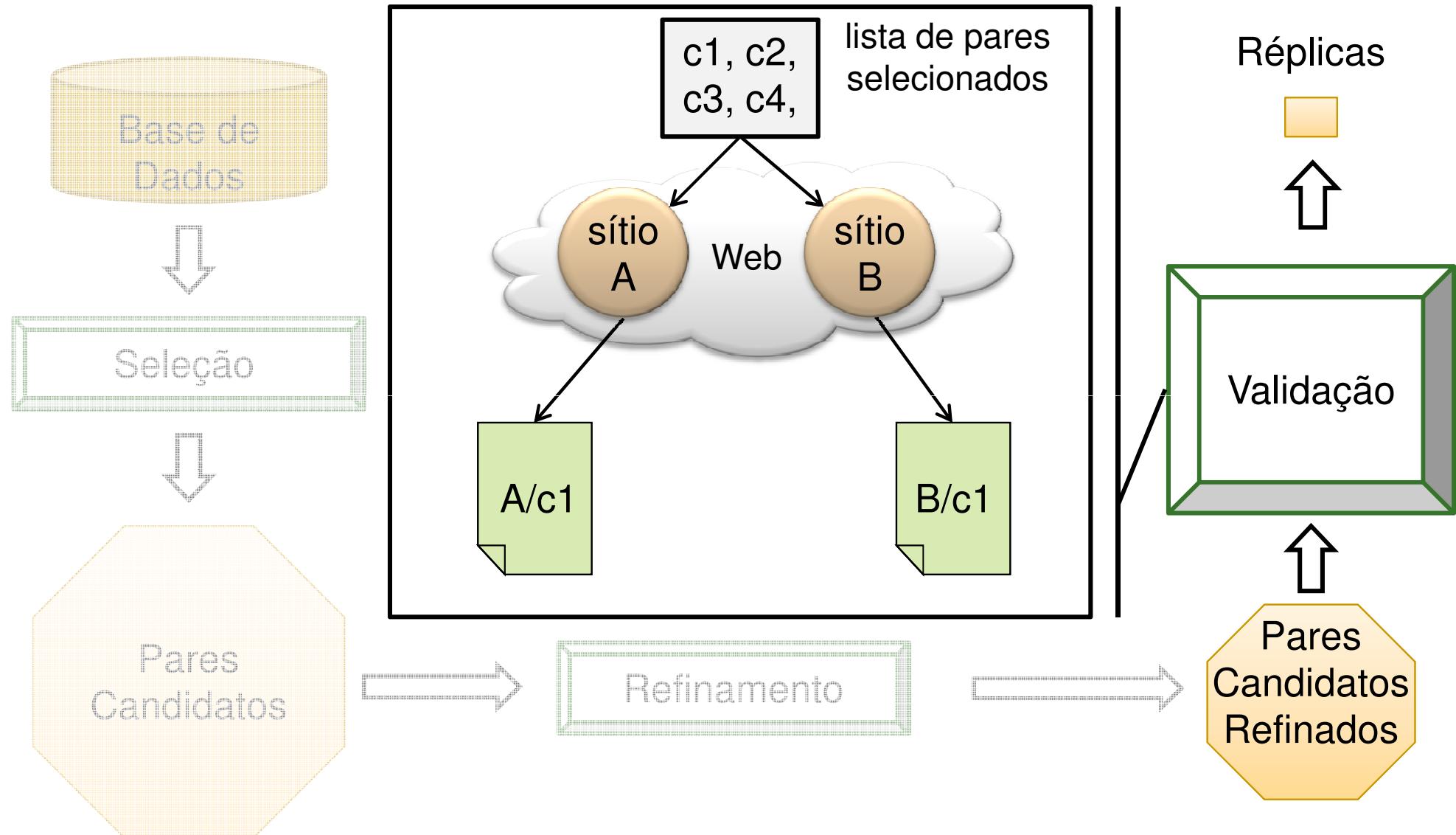
2^a Fase: Aplicação do Modelo Aprendido



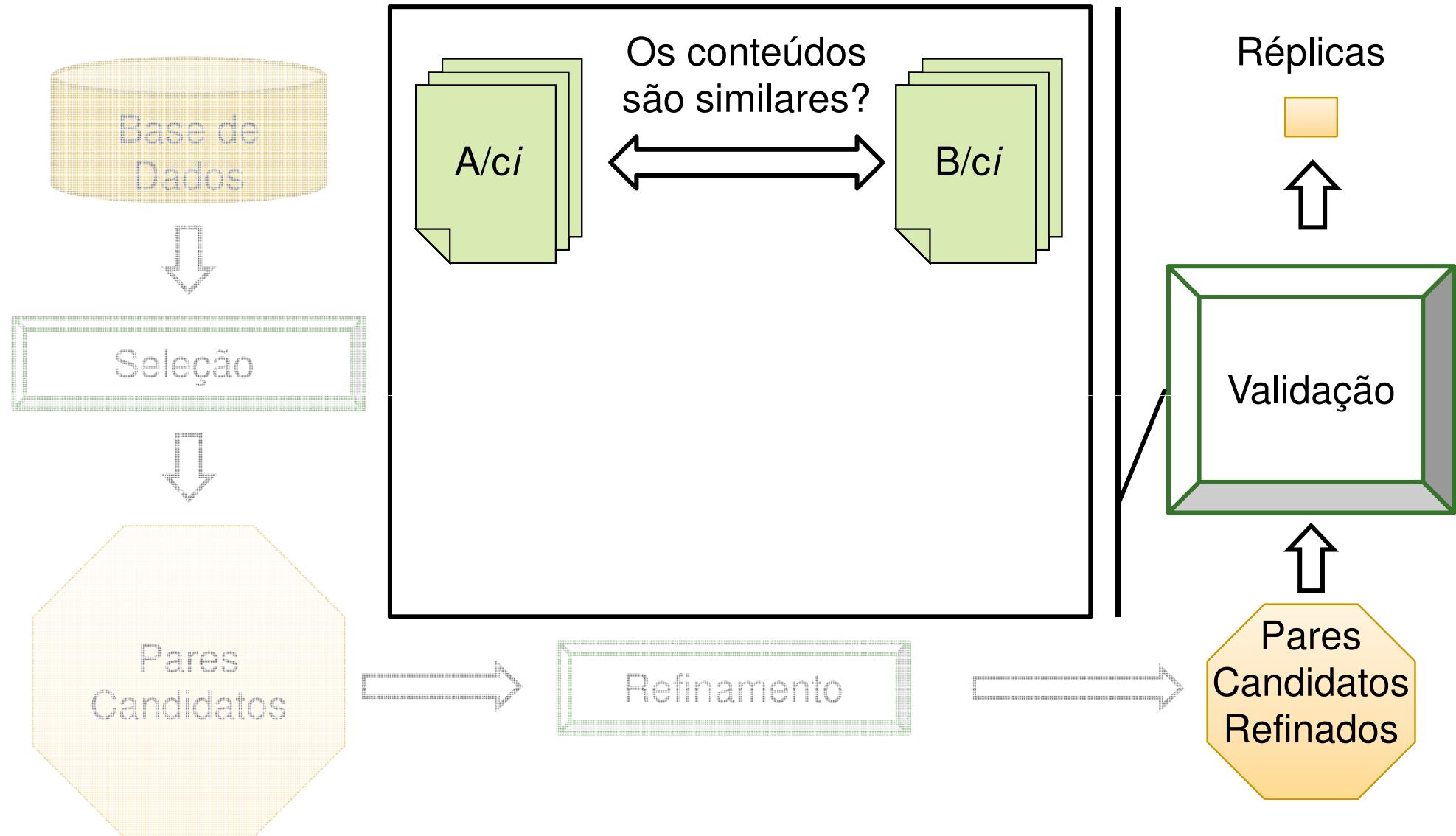
3^a Fase: Seleção Aleatória de Caminhos



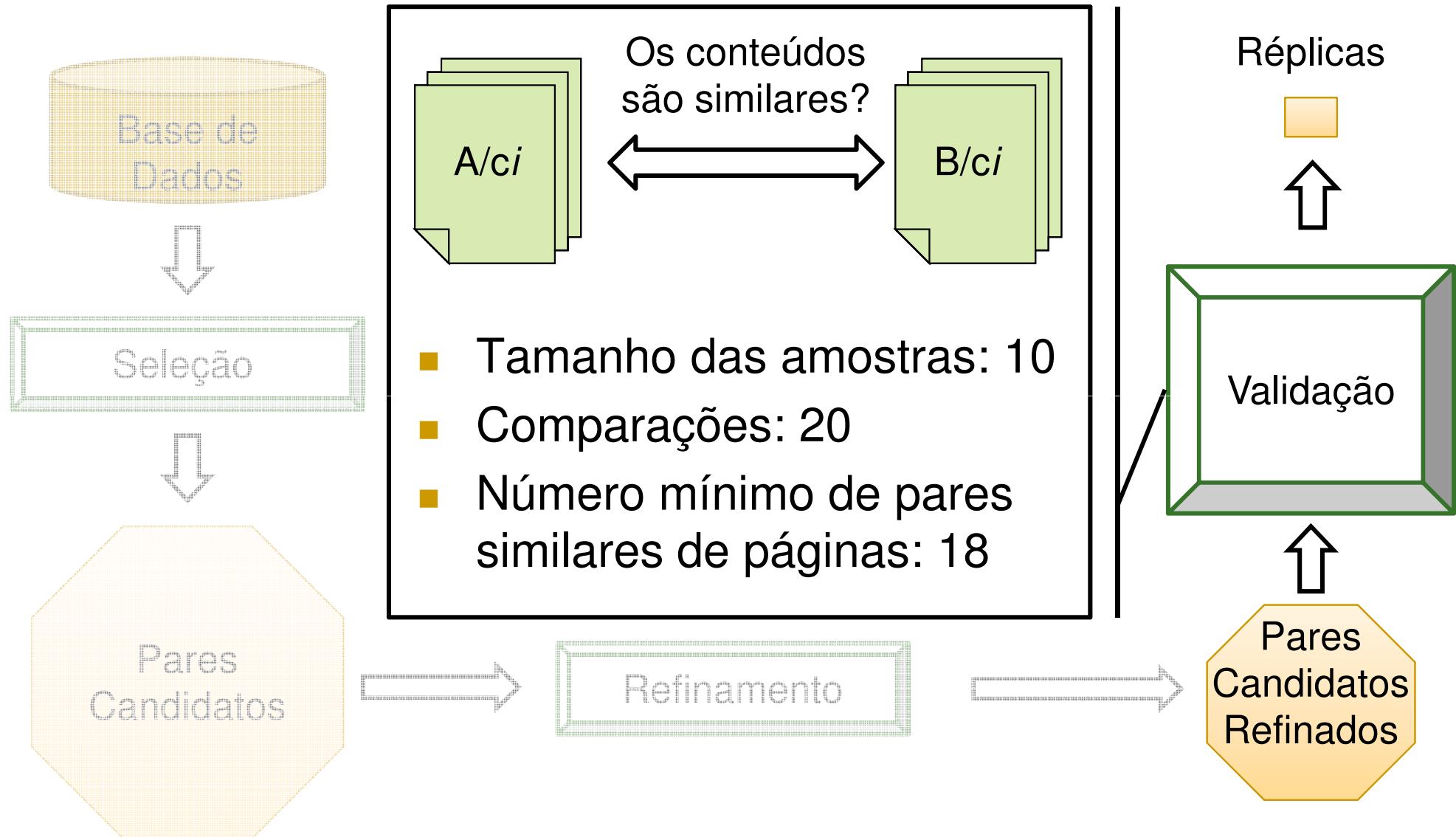
3^a Fase: Coleta da Lista de Caminhos



3^a Fase: Comparação Entre Conteúdos



3^a 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 (1^ª fase DREAM)
- O desempenho do *Baseline* não foi afetado
 - Recupera um subconjunto do conjunto selecionado pela 1^ª fase do DREAM

Geração do Gabarito

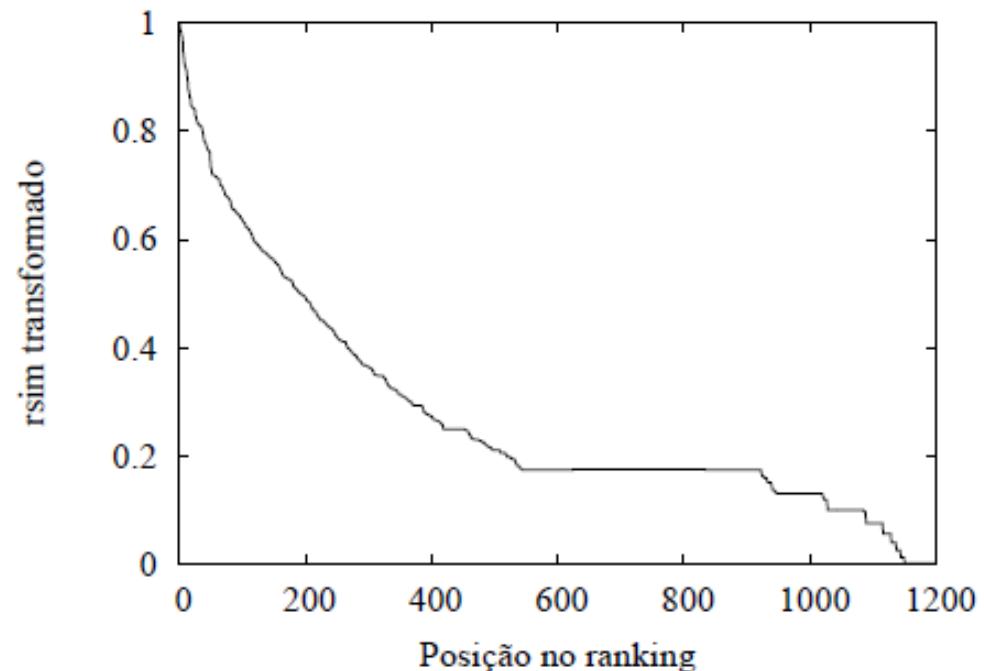
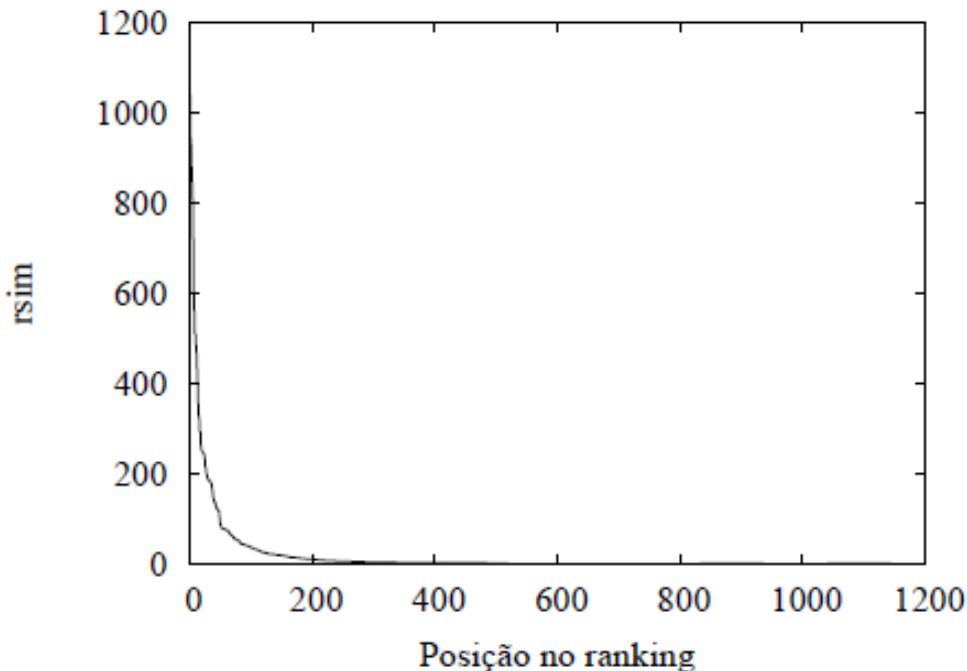
- Pares retornados pela 1^a Fase: **8.906.742**
- O custo do processo de validação ainda é alto
- Tamanho escolhido para o gabarito: **10.000**
 - Amostragem aleatória dos 8.906.742
 - Vale lembrar que a redução no tamanho do gabarito não interfere em nossos objetivos
- Pares de réplicas encontradas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

Classes de sítios

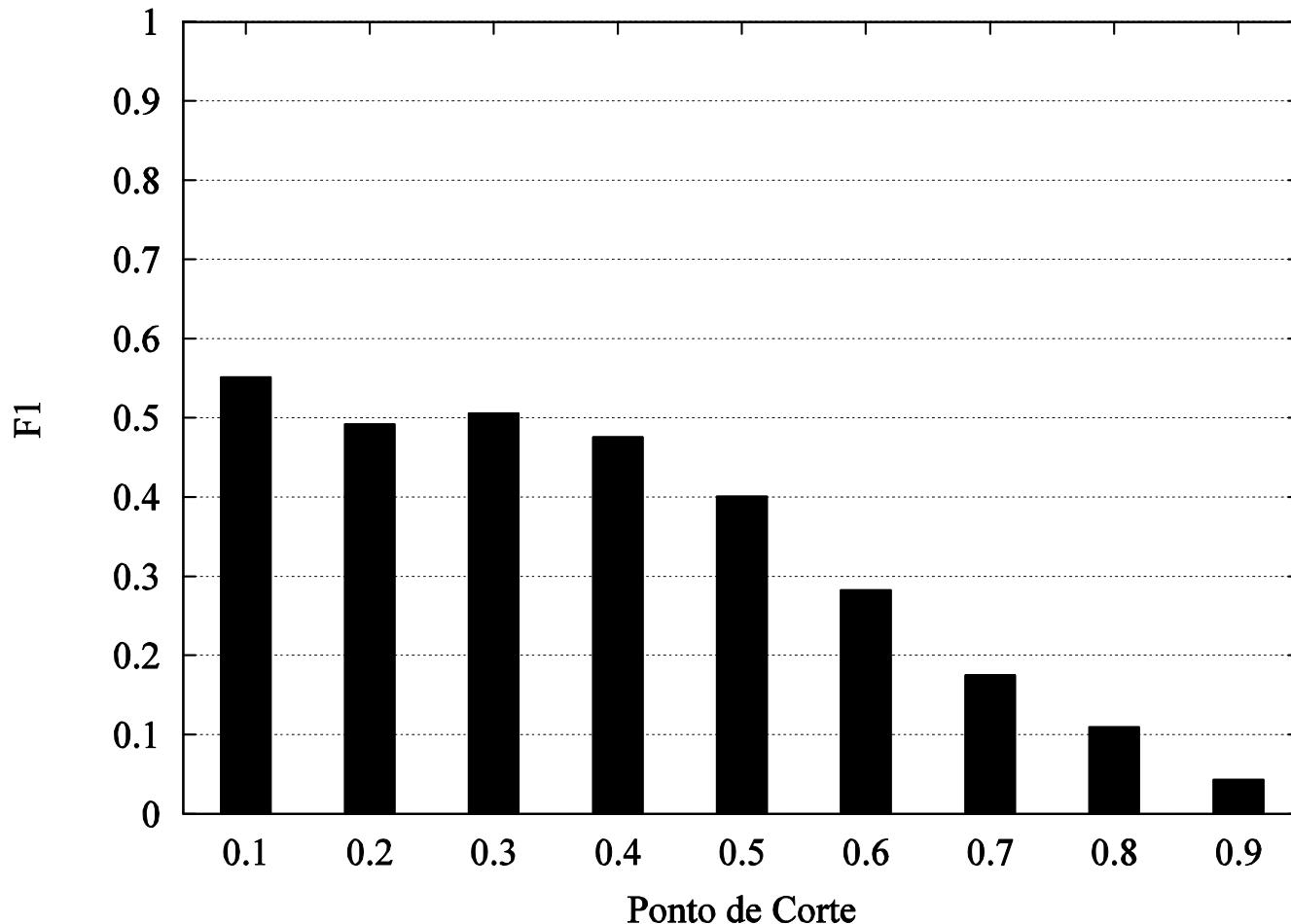
- Conteúdo volátil
 - Sítios de ofertas
 - www.ksesporte.com.br e www.materialdenatacao.com.br
- Caminhos trocados
 - Sítios hospedados no mesmo servidor
 - arte.centralblogs.com.br e downloads.centralblogs.com.br
- Conteúdo regional
 - Jornais regionais
 - ac.noticiannahora.com.br e ms.noticiannahora.com.br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

Ajuste de Parâmetros

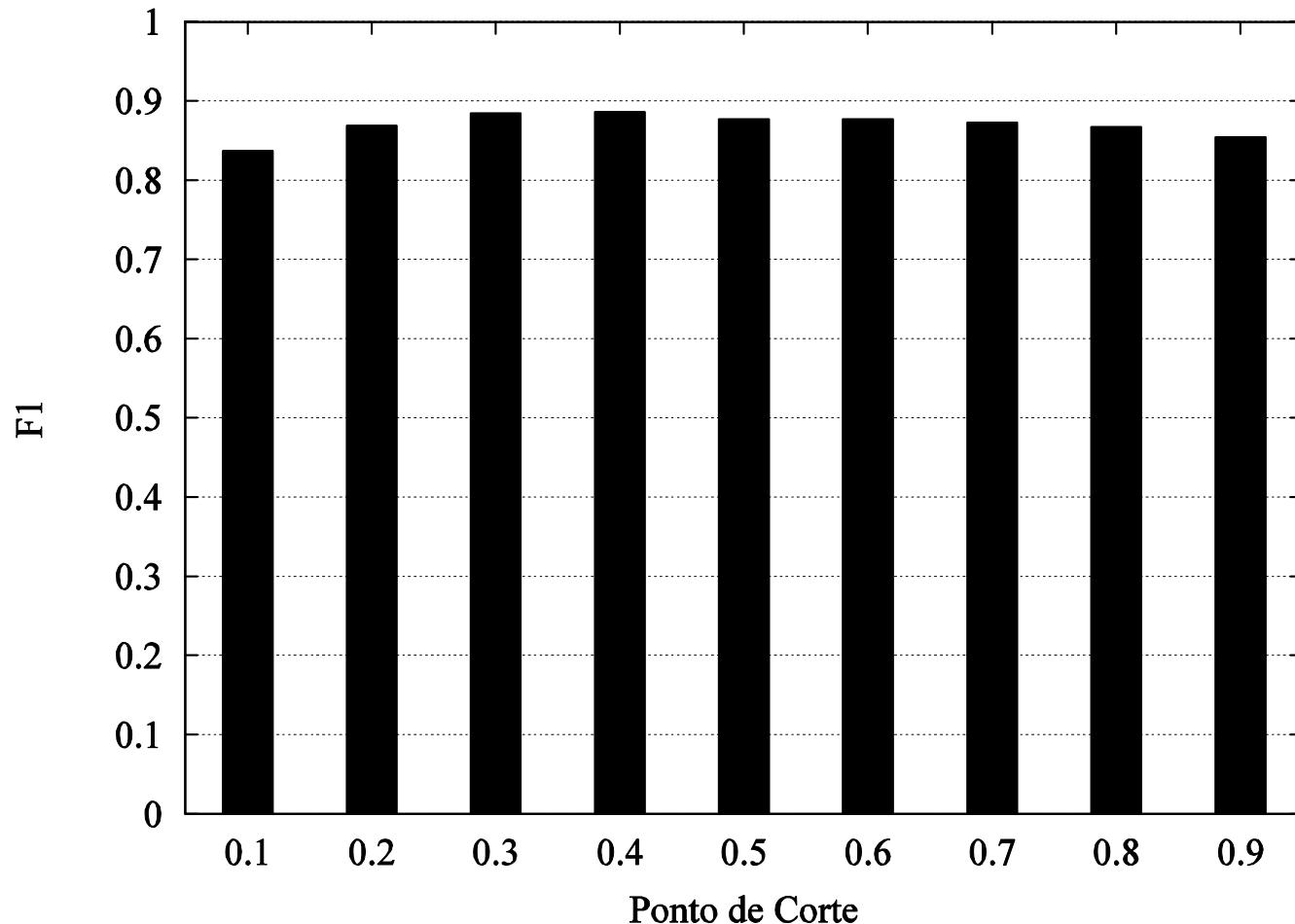
- Os algoritmos DREAM-X retornam uma probabilidade de o par em questão ser réplica
- NormPaths retorna uma pontuação: $rsim$
- \log e normalização foram aplicados sobre $rsim$



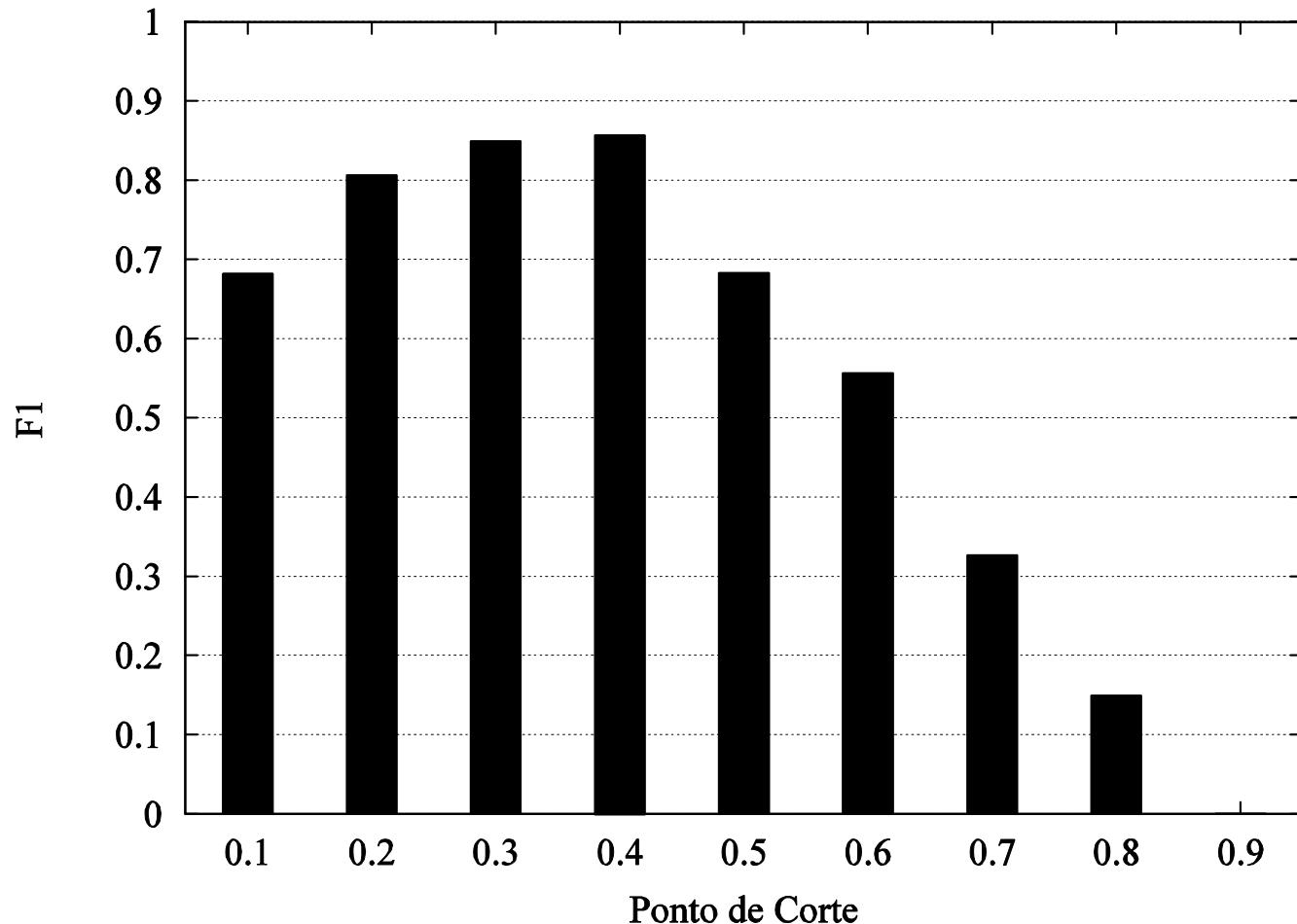
Níveis de F1 do Algoritmo NormPaths



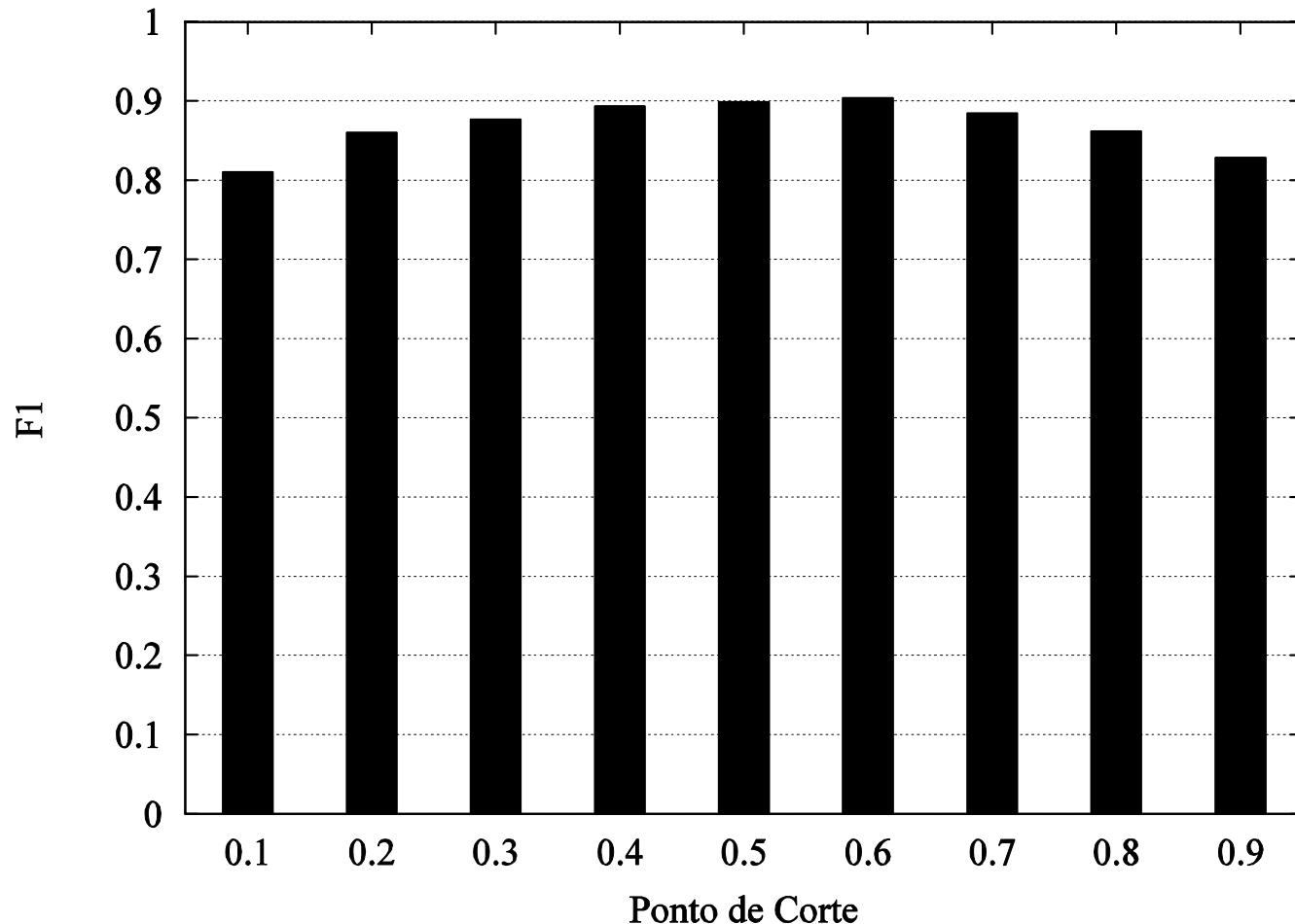
Níveis de F1 do Algoritmo DREAM-DT



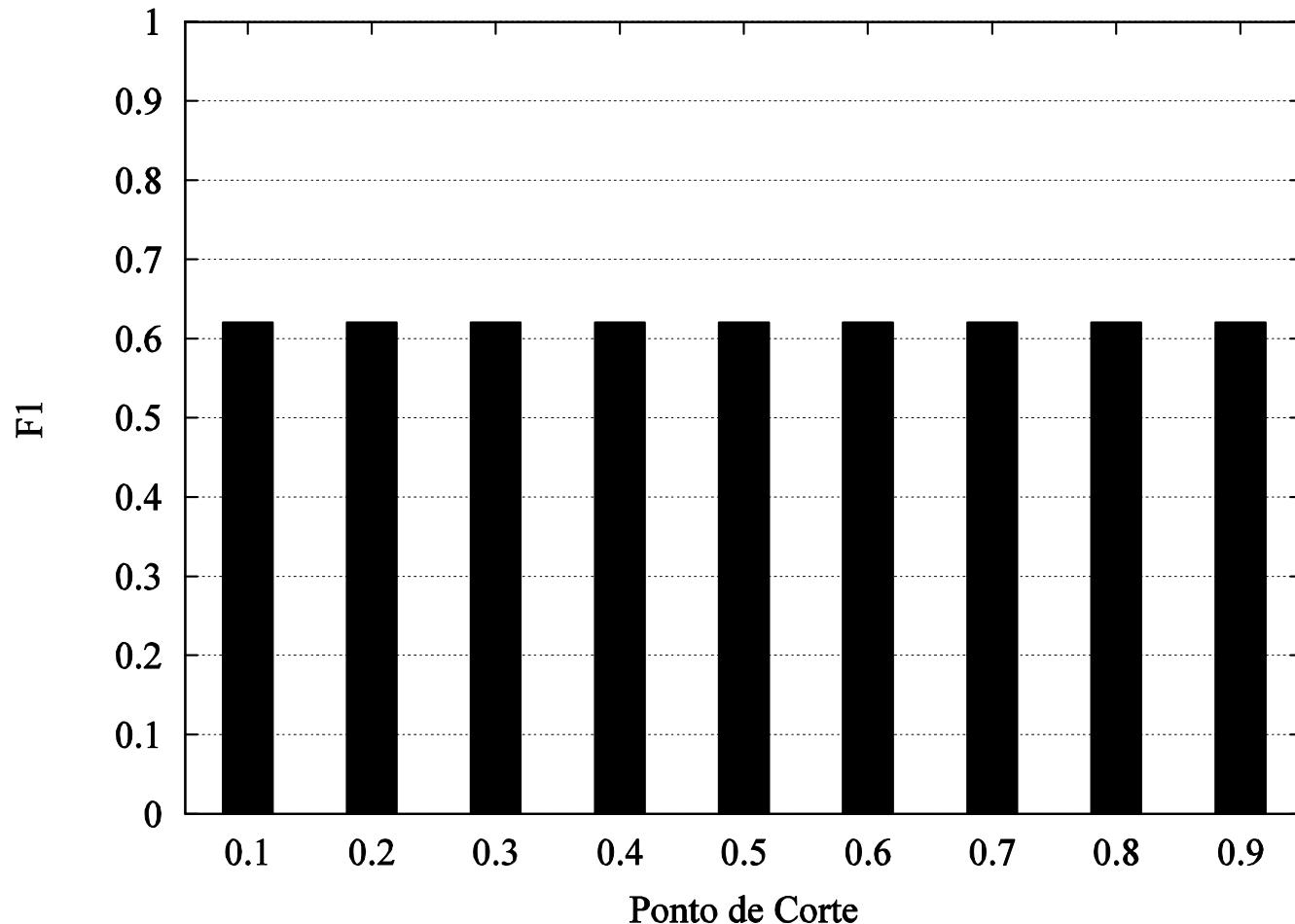
Níveis de F1 do Algoritmo DREAM-LAC



Níveis de F1 do Algoritmo DREAM-RF

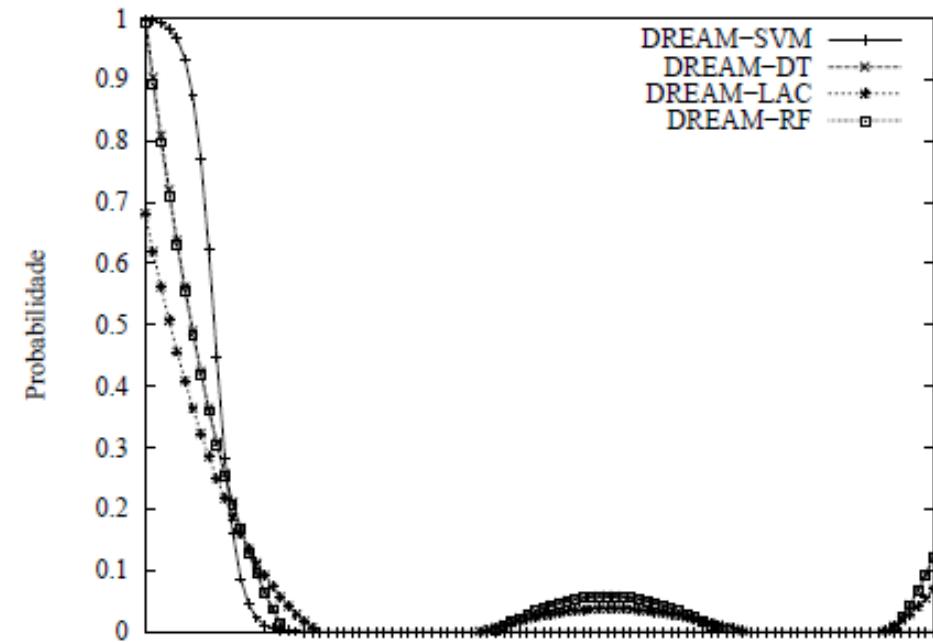
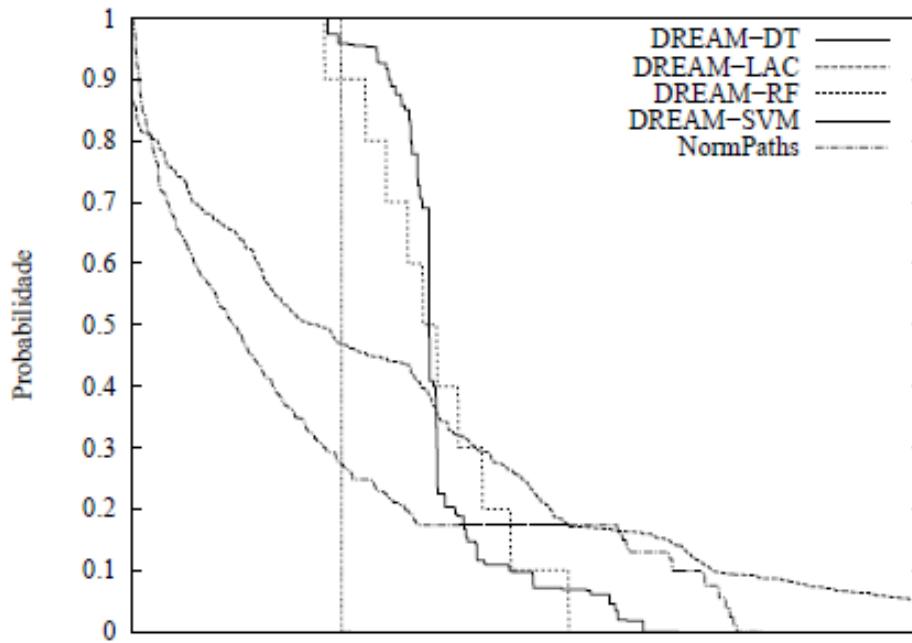


Níveis de F1 do Algoritmo DREAM-SVM



Distribuições de Probabilidade

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



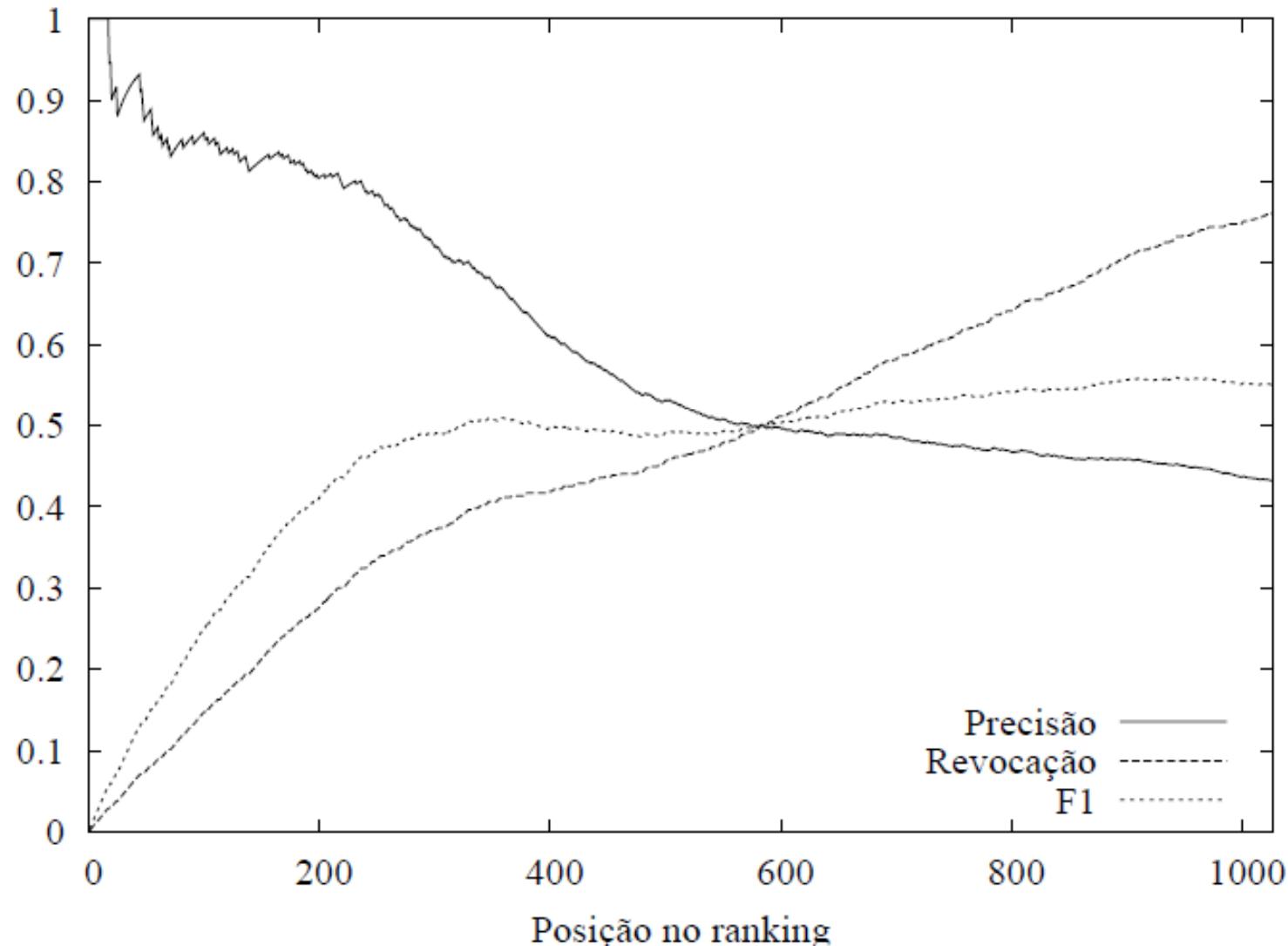
Resultados

■ Ajuste de parâmetros

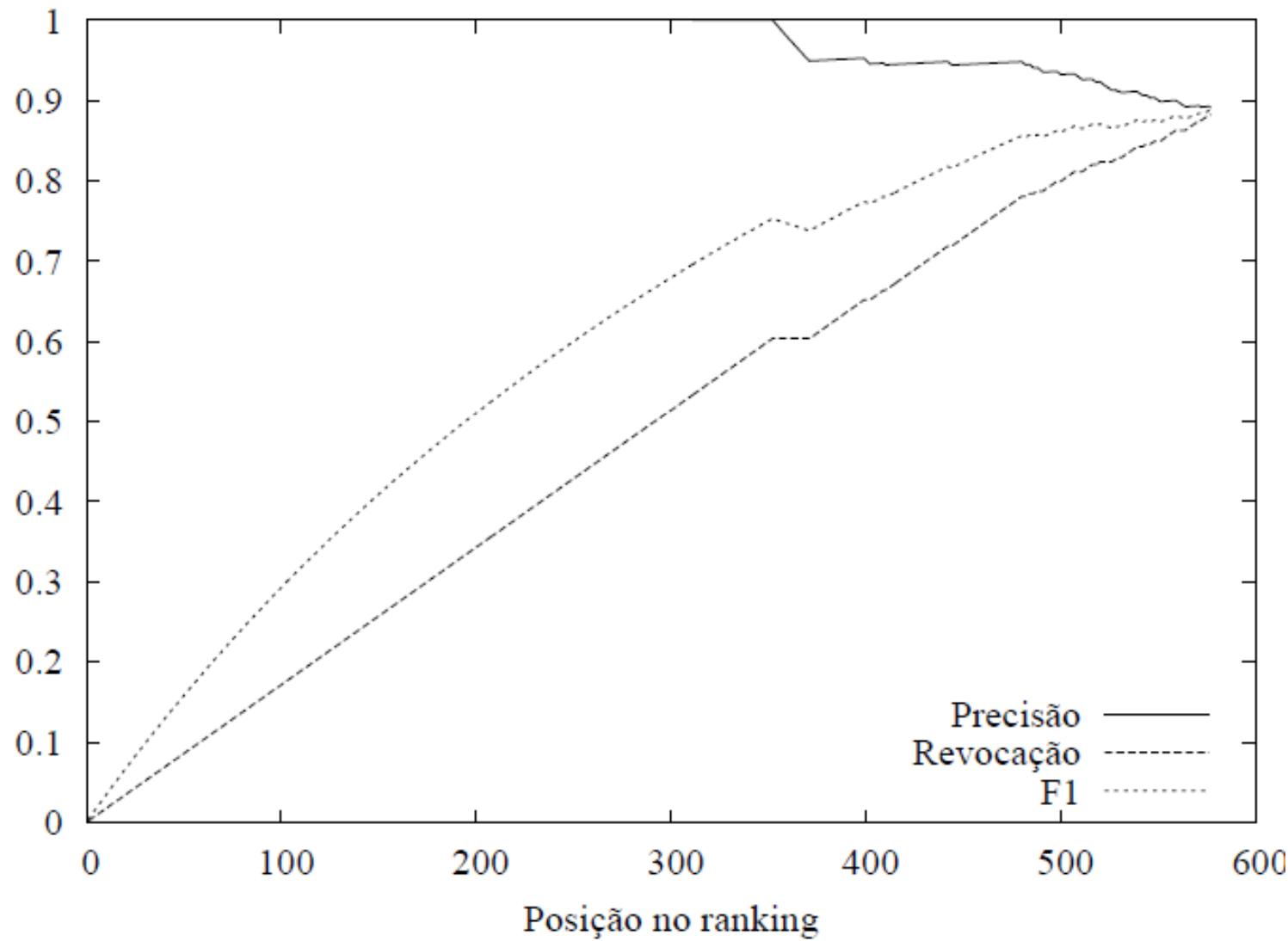
Algoritmo	Ponto de Corte
NormPaths	0,1
DREAM-DT	0,4
DREAM-LAC	0,4
DREAM-RF	0,6
DREAM-SVM	-

- 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

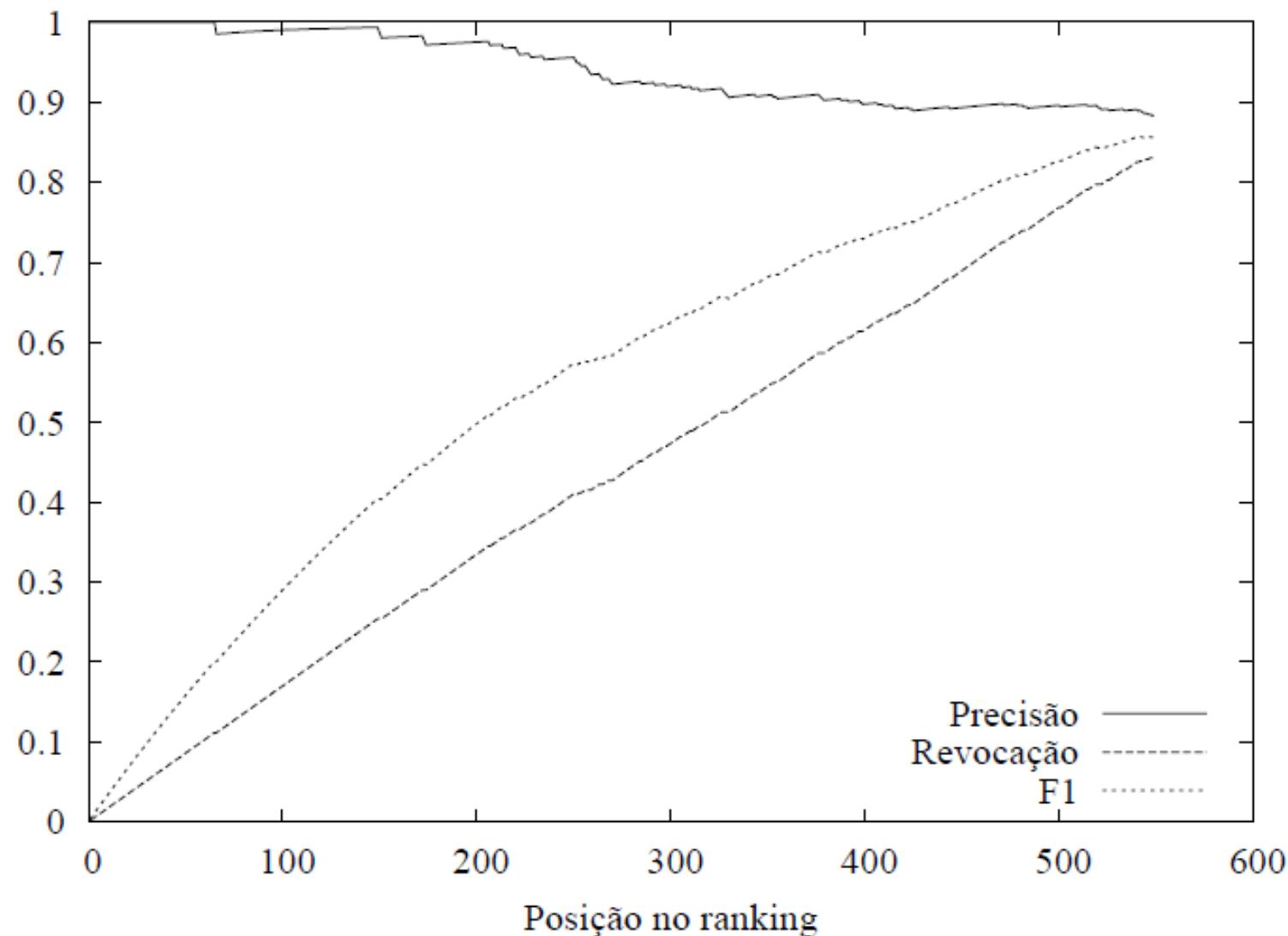
Desempenho do Algoritmo NormPaths



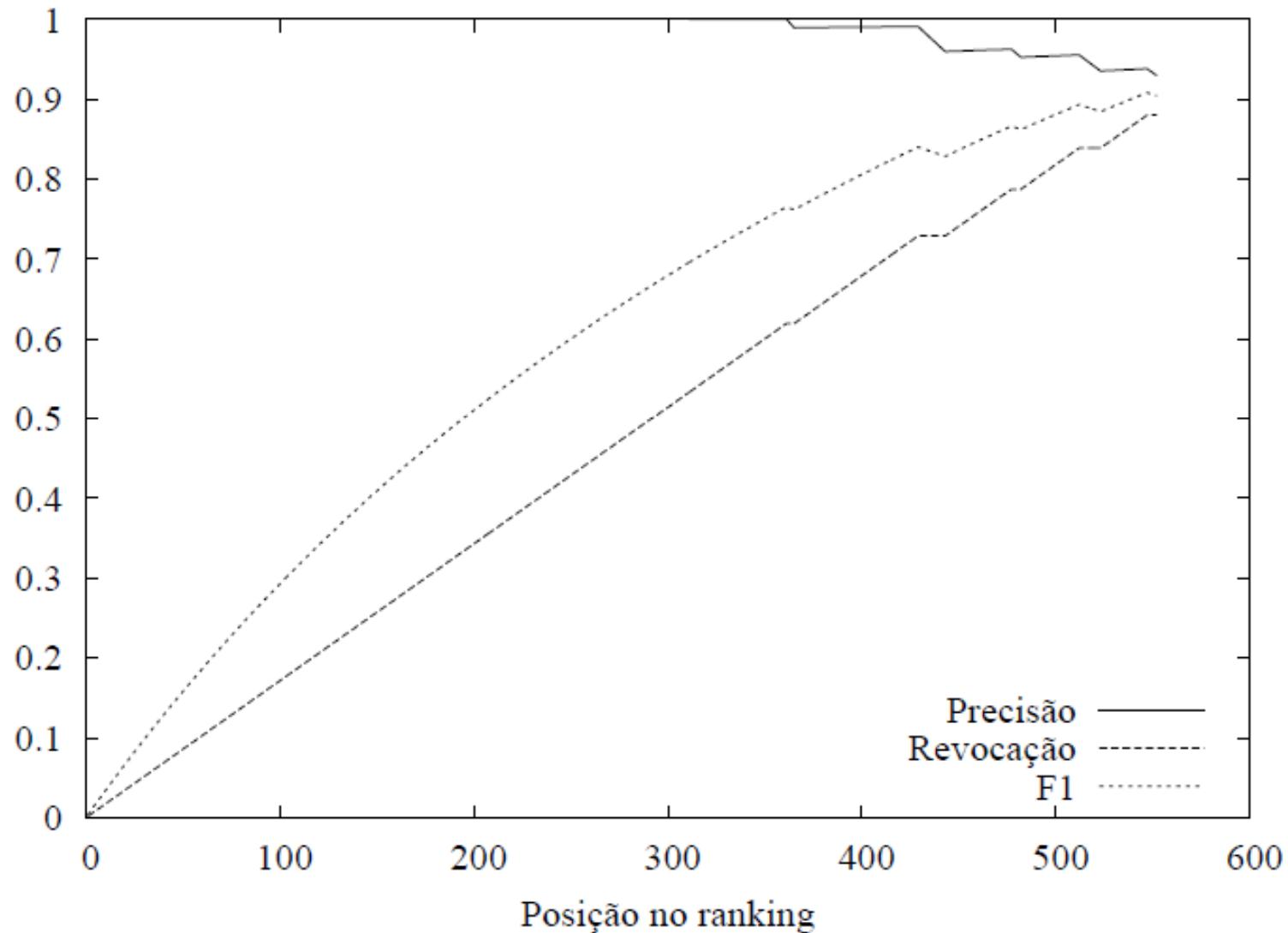
Desempenho do Algoritmo DREAM-DT



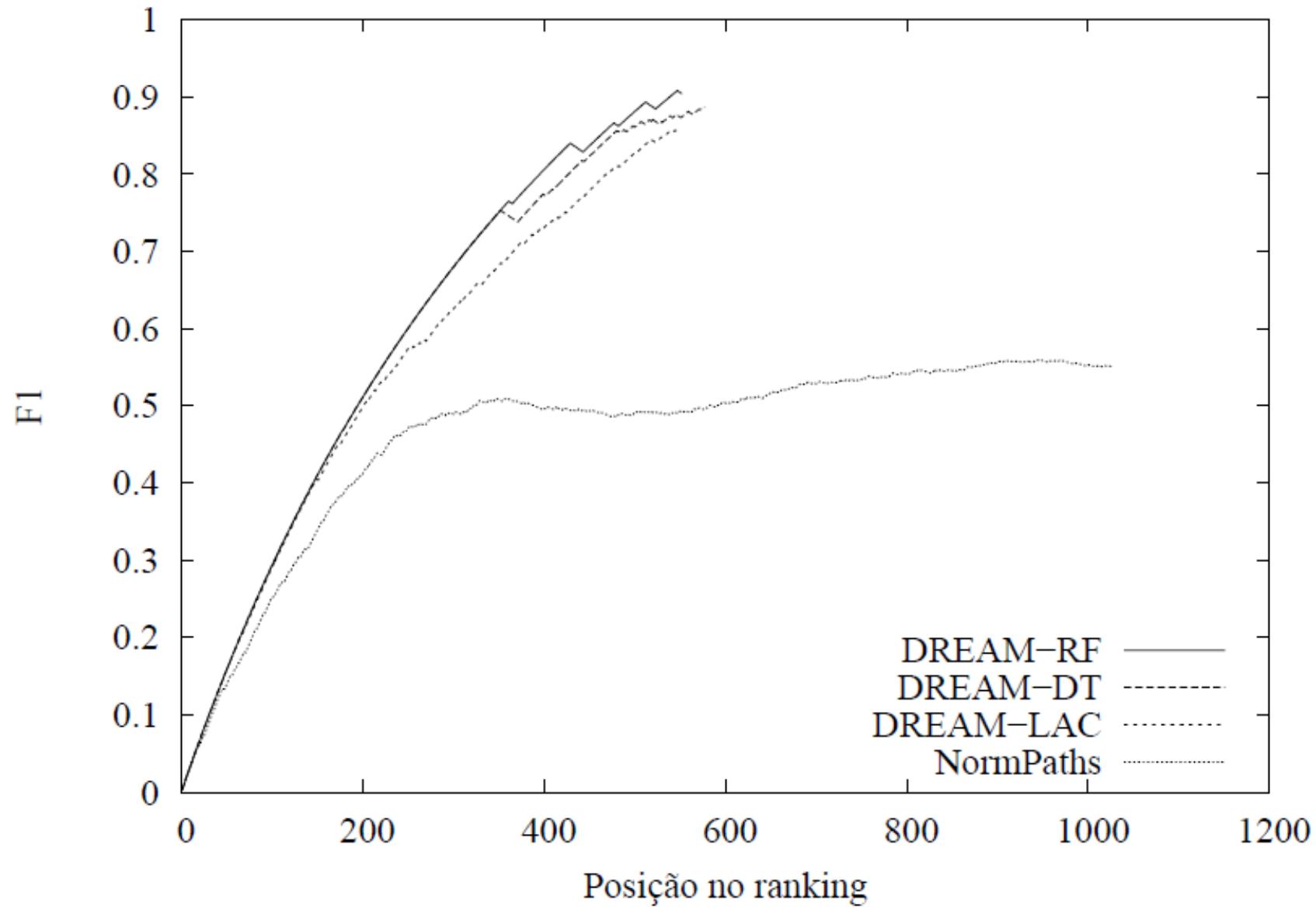
Desempenho do Algoritmo DREAM-LAC



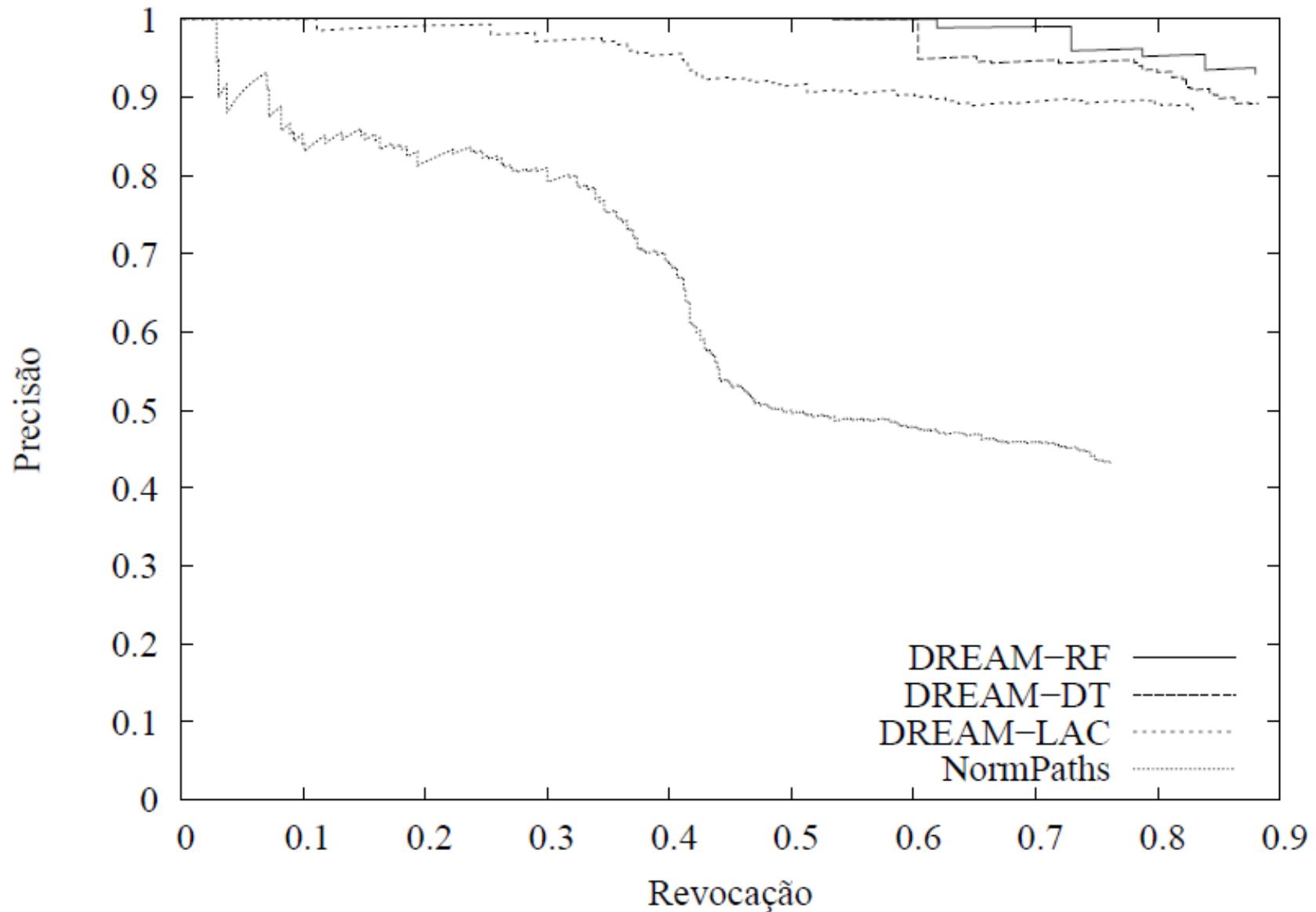
Desempenho do Algoritmo DREAM-RF



Comparação: Curvas de F1



Comparação: Precisão x Revocação



Resultados Finais

Algoritmo	Melhor Ponto		Revocação	F1	Ganho
	de Corte	Precisão			
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?

