Detecção de Réplicas de Sítios Web Usando Aprendizado Semi-supervisionado baseado em Maximização de Expectativas

Cristiano Rodrigues de Carvalho
Orientador: Nivio Ziviani
Co-orientador: Adriano Veloso

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

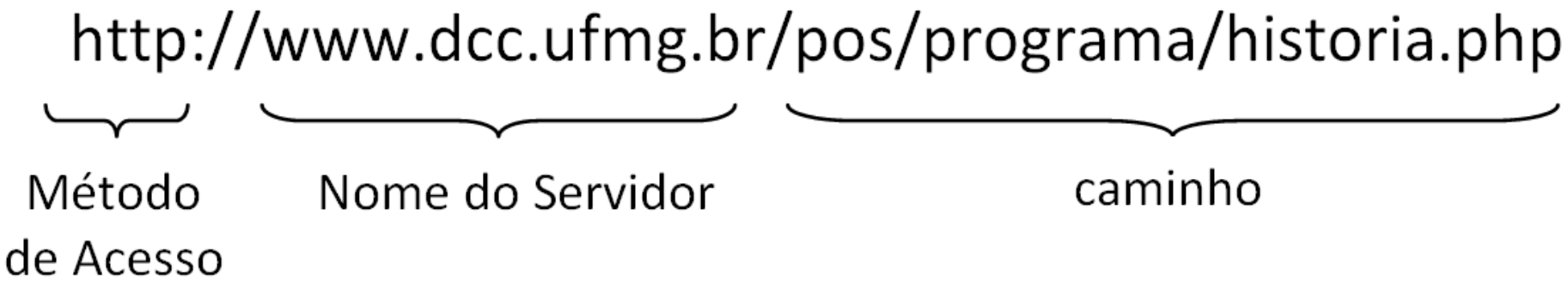
19 de Setembro de 2014

Problema

- 29% de conteúdo duplicado na Web
- Duplicação intra-sítios e inter-sítios
- Sistemas de busca armazenam cópias da Web
- Desperdício de recursos, anomalias no ranking, respostas repetidas

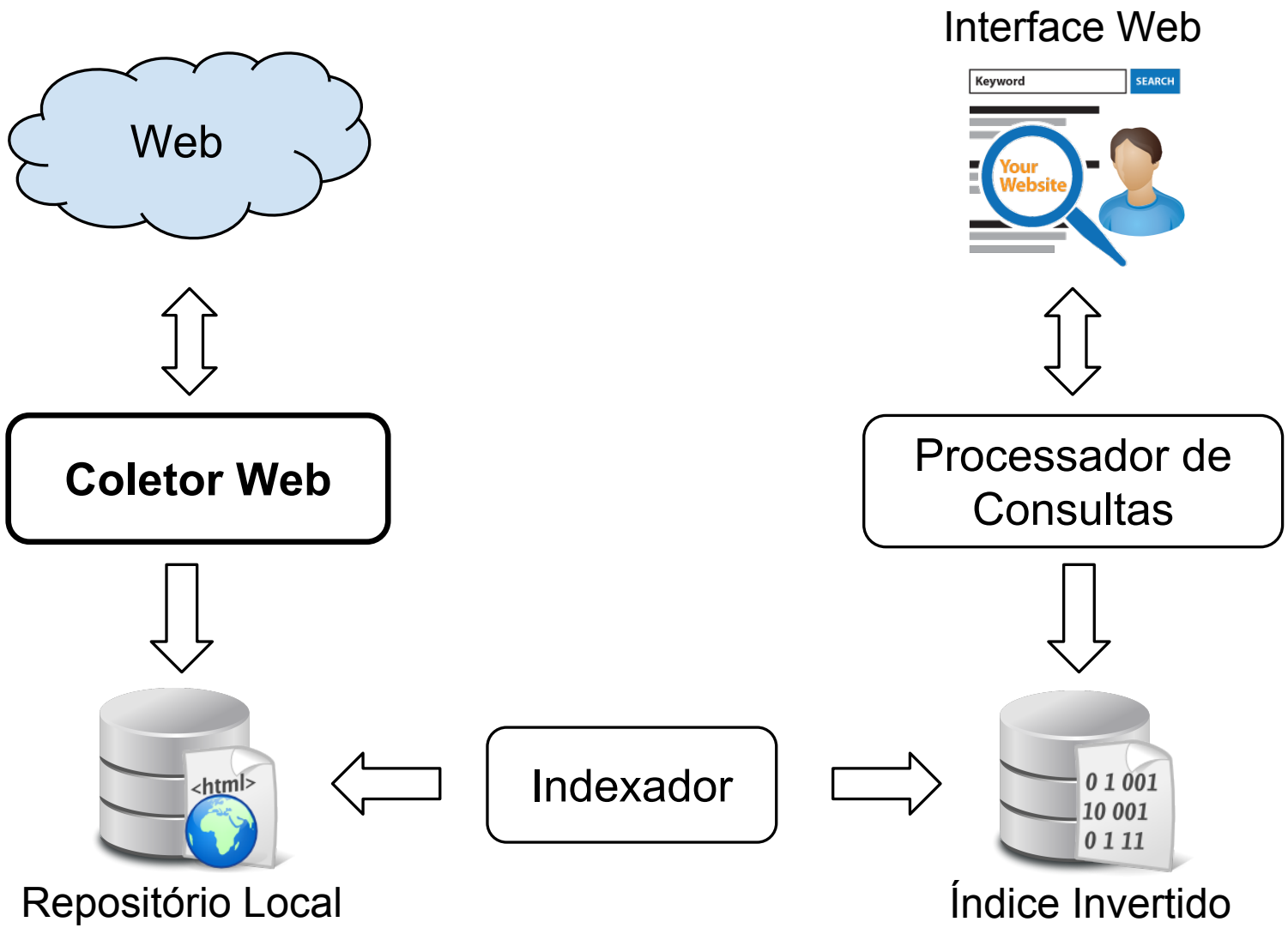
Reduzir o número de duplicatas através da detecção réplicas de sítios web

Sítios Web

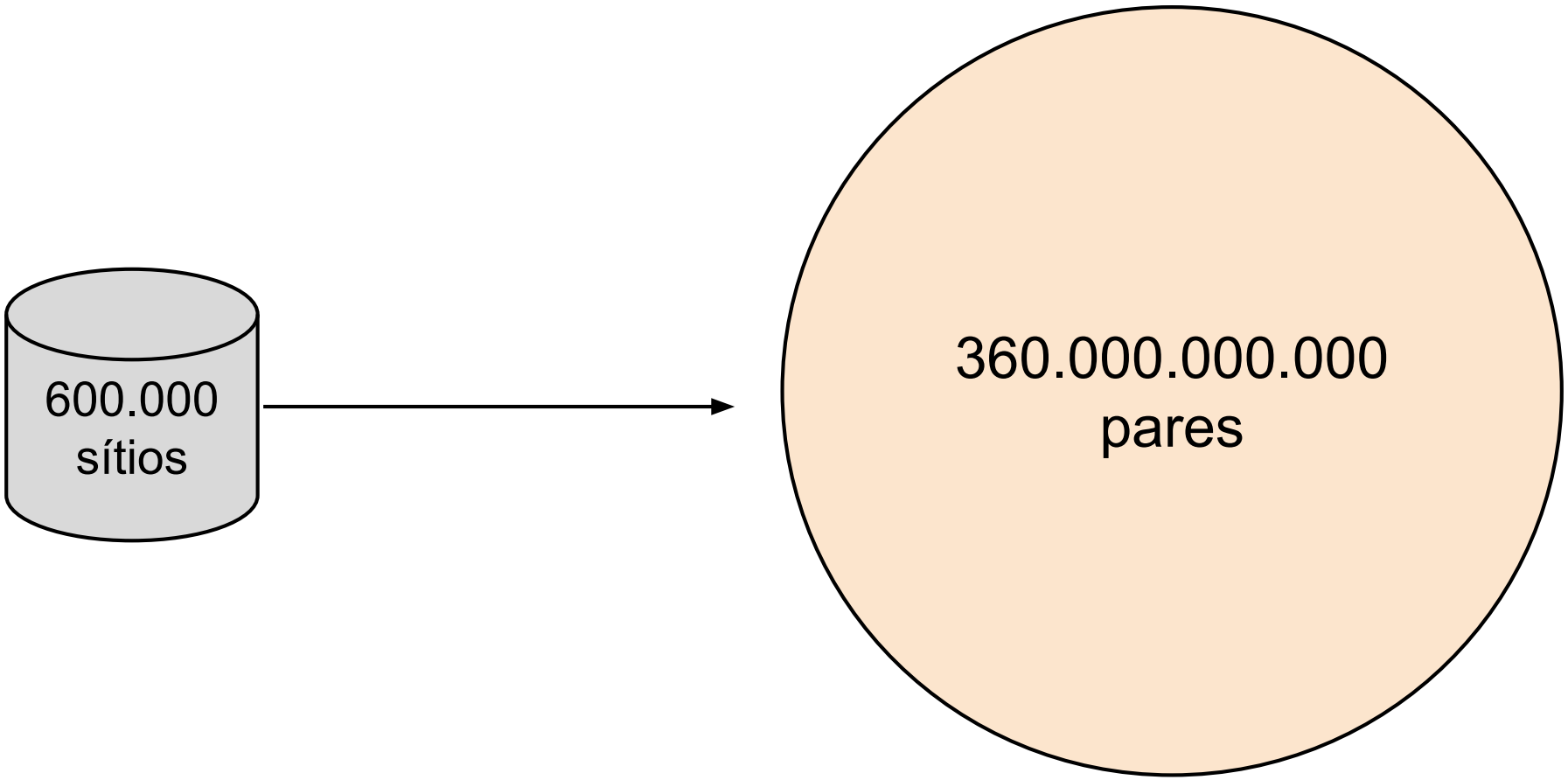


- Sítios: páginas web que compartilham o mesmo nome de servidor

Máquinas de Busca: Componentes

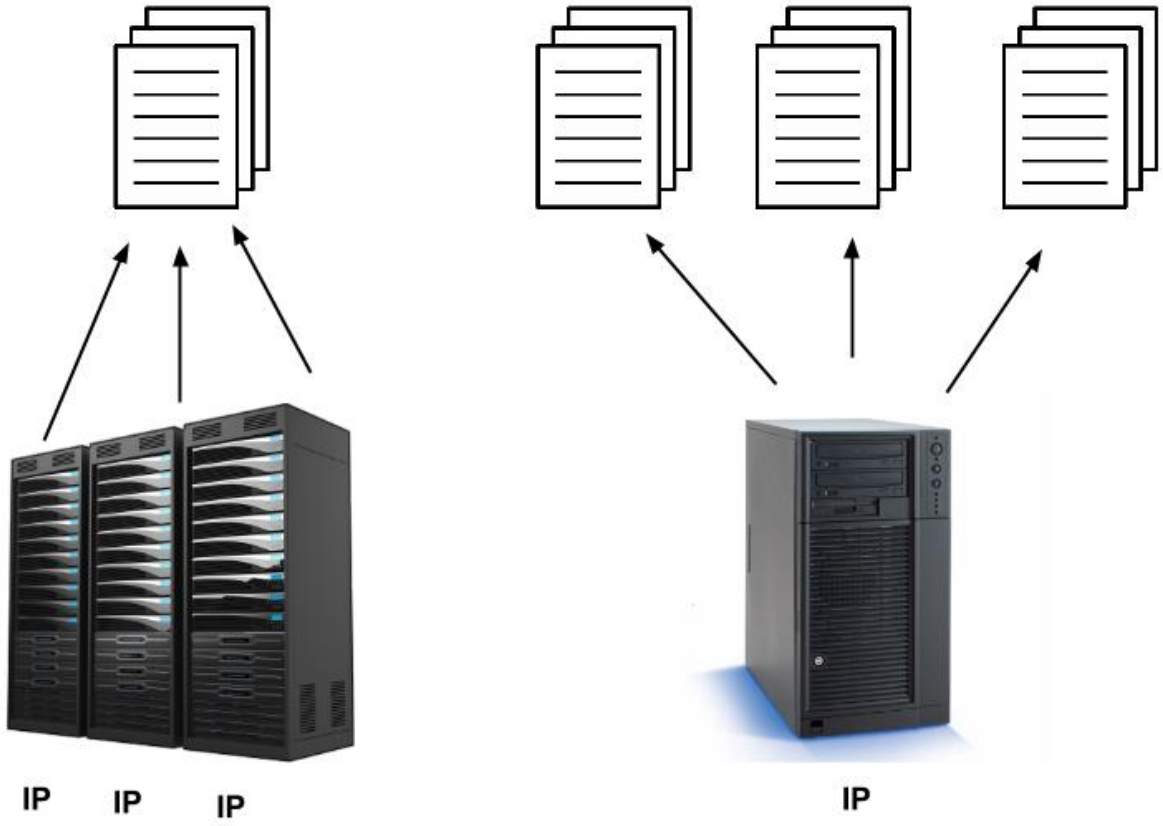


Desafios



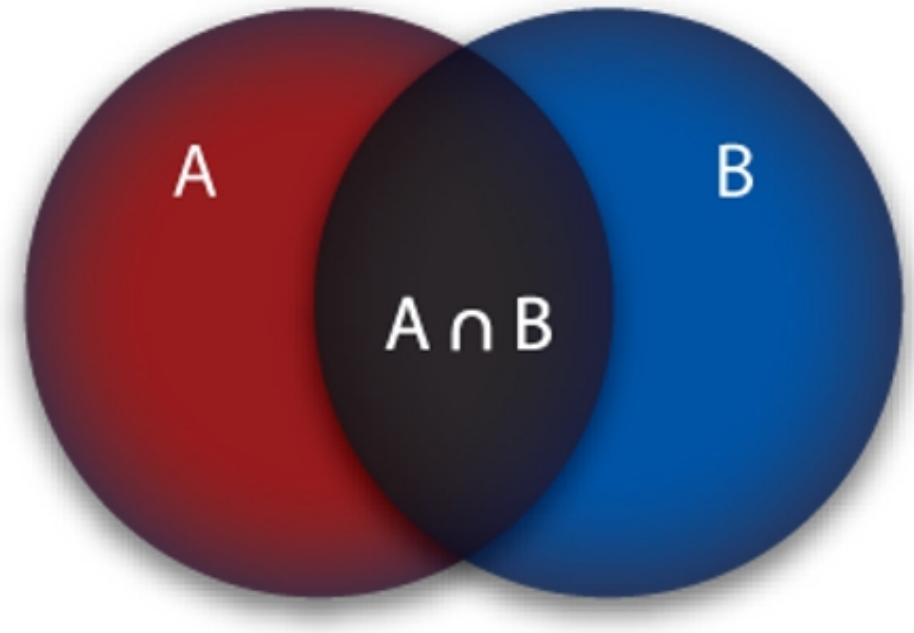
- Problema quadrático

Desafios



- Vários IPs para um único sítio e vários sítios para um único IP.

Desafios



- Pequena interseção de páginas conhecidas de sítios replicados
- Sítios muito similares que não são réplicas (cifras)

Solução Proposta

- Algoritmo baseado em aprendizado de máquina
- Duas fases:
  - Seleção de sítios candidatos a réplica
  - Classificação de pares candidatos
- Necessidade de treino rotulado
  - Abordagem semi-supervisionada para aquisição de exemplos

Trabalhos Relacionados - Inter-sitios

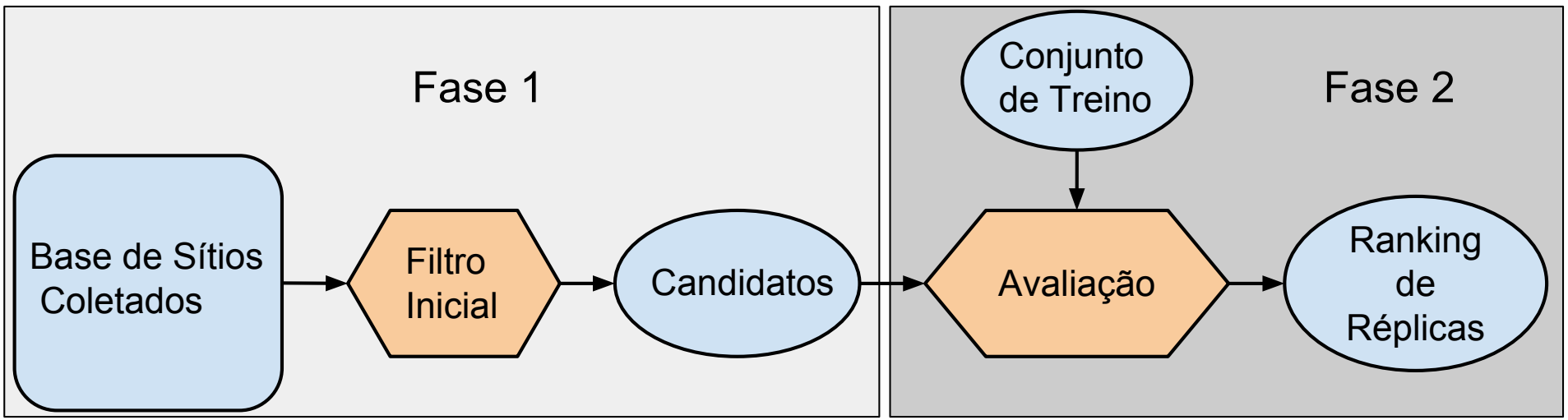
- Estudo sobre replicação na Web
  [Bharat et al.,1999]
- Heurísticas para detecção de réplicas
  [Bharat et al.,2000]
- Uso eficiente de informações de conteúdo
  NormPaths [da Costa Carvalho et al.,2007]
  - Estado-da-arte e baseline

Trabalhos Relacionados - Intra-sitios

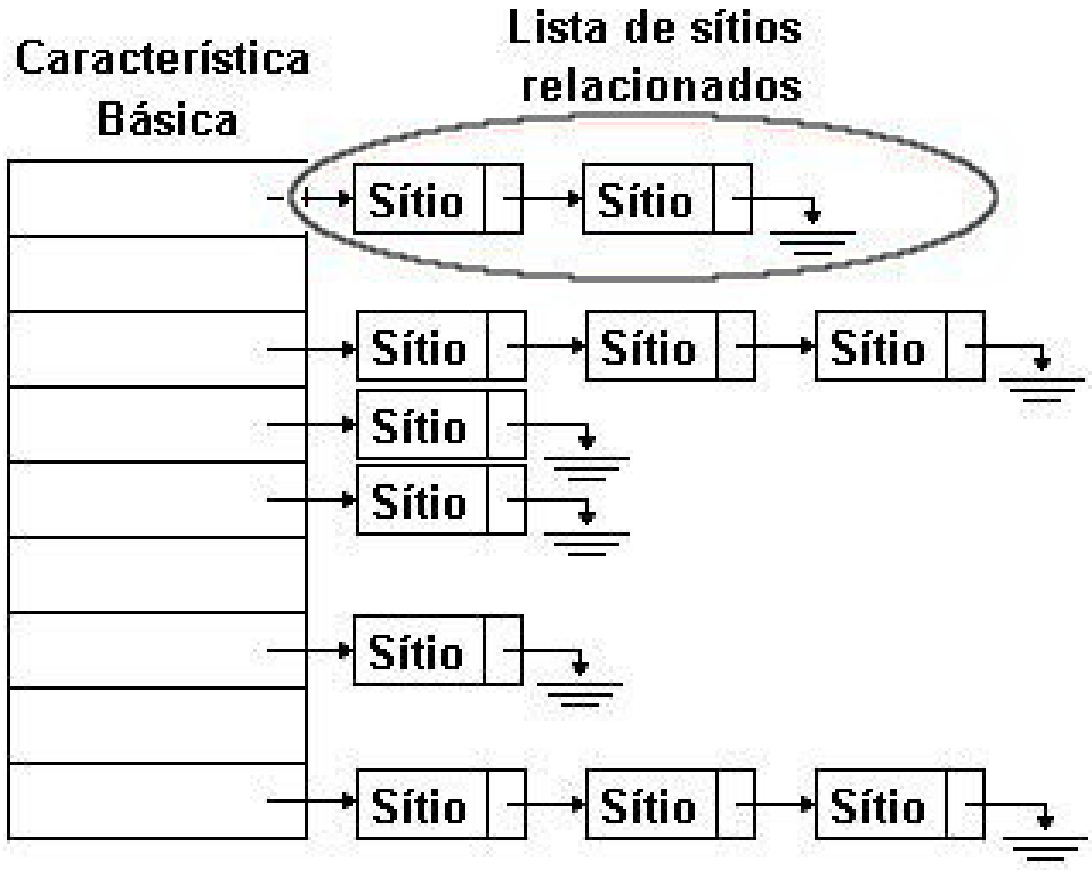
- DUST (Different URLs with Similar Text)
  [Bar-Yossef et al., 2009]
  [Koppula et al., 2010]
  [Rodrigues et al., 2013]
  - Detecção de conteúdo similar em URLs distintas
  - Diversos trabalhos recentes

Metodologia

O Algoritmo Proposto



Filtro de Candidatos



- Caminho da URL
- Assinatura do Conteúdo

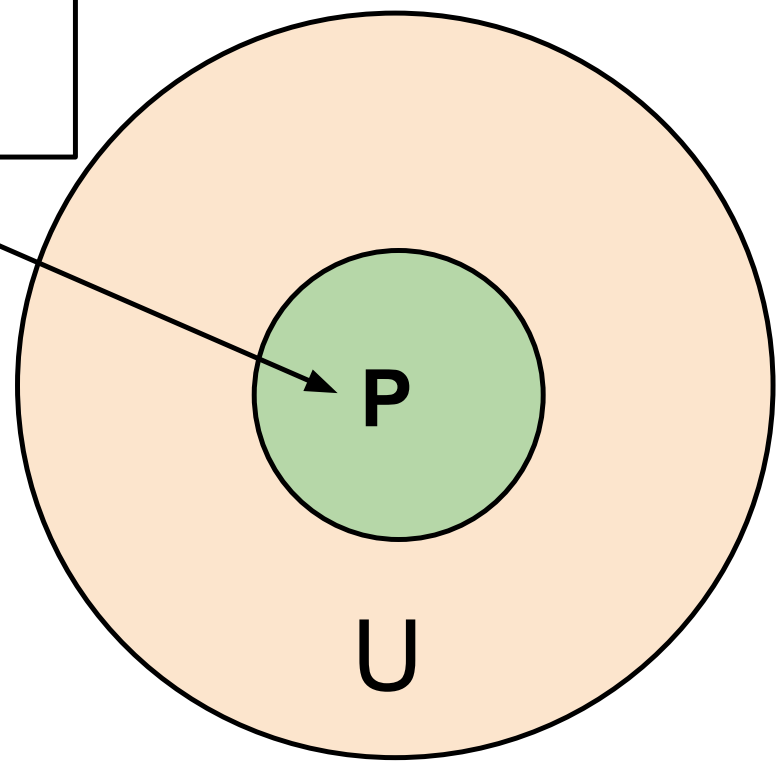
Avaliação de Candidatos

- Características de refinamento
  - Distância de Edição (*ndist*)
  - Correspondência de Nomes de Servidor (*nmatch*)
  - Quatro Octetos (*ip4*)
  - Três Octetos (*ip3*)
  - Correspondência entre Caminhos Completos (*fullpath*)
- Algoritmo de Classificação LAC
  `[Veloso & Meira Jr., 2011]`

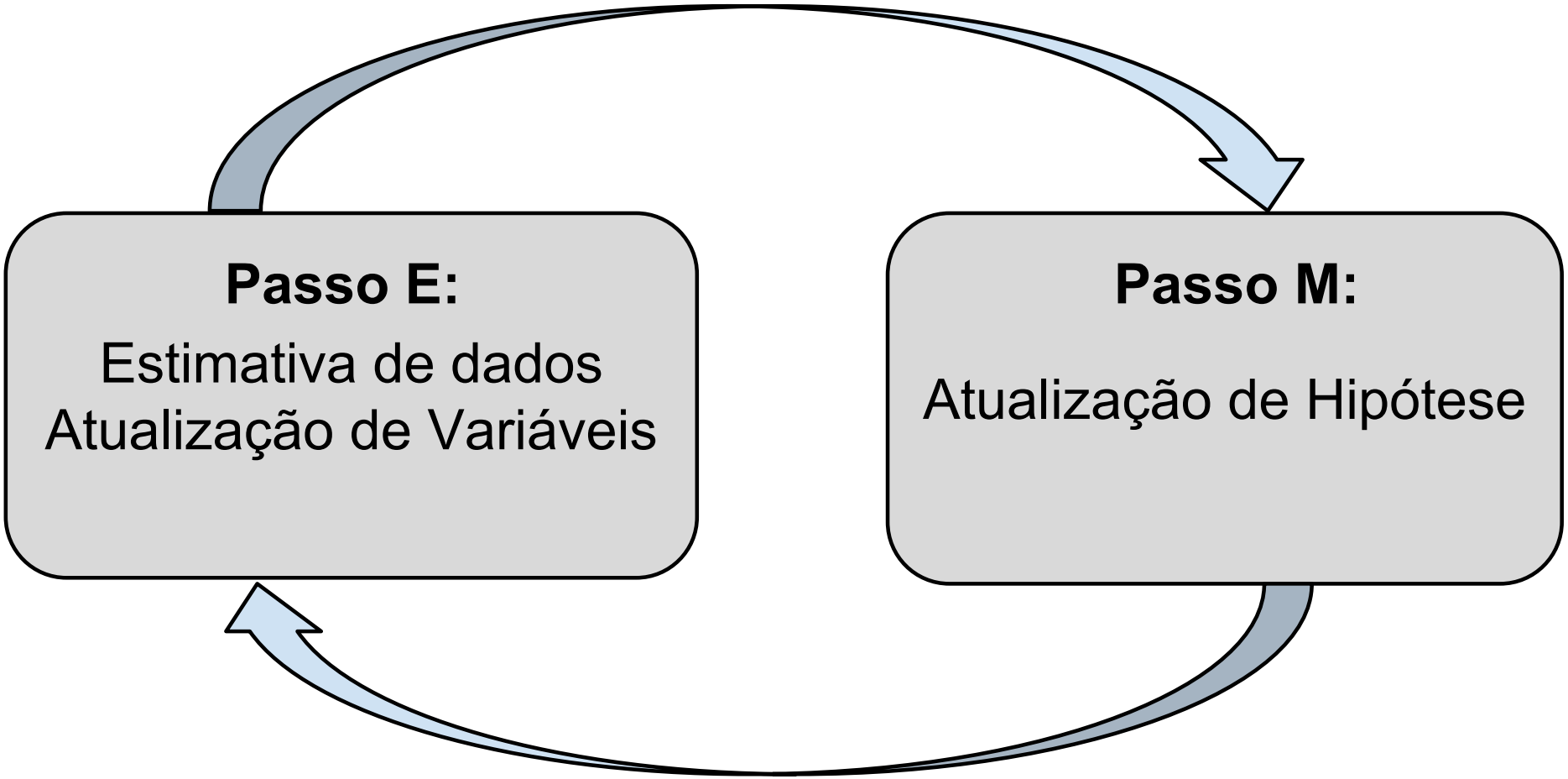
Modelos de Treino: PU

www.exemplo.com.br == exemplo.com.br

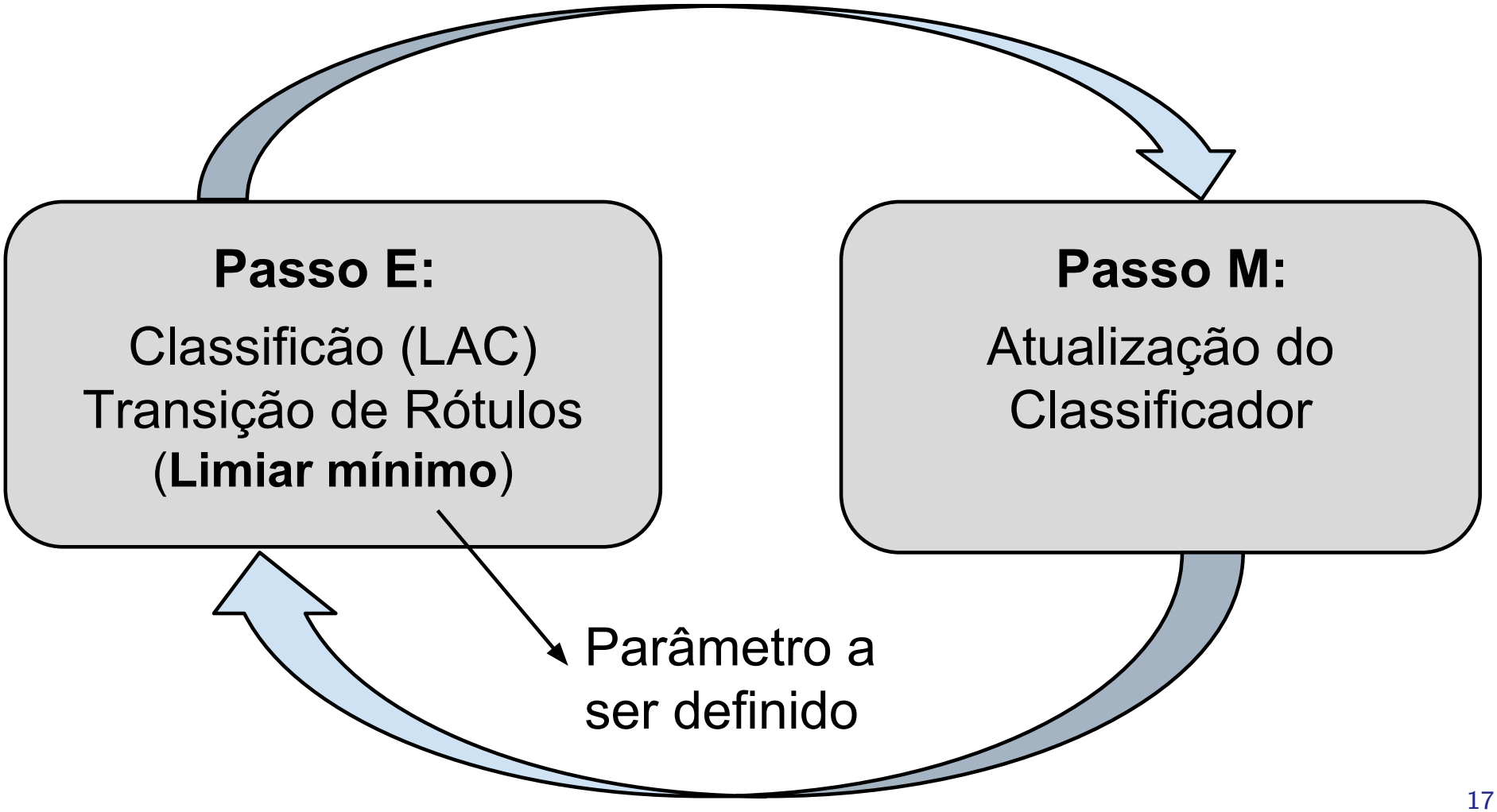
exemplo.**gov.br** == exemplo**.br**



Modelos de Treino: EM



Modelos de Treino: EM



Modelos de Treino: Escolha de Limiar

- Limiar global único
  - Diferentes valores de teste (0.1 a 0.9)
- Limiar local ótimo
  - Projeção de treino
  - Abordagem de entropia mínima
    [Davis et al., 2012]

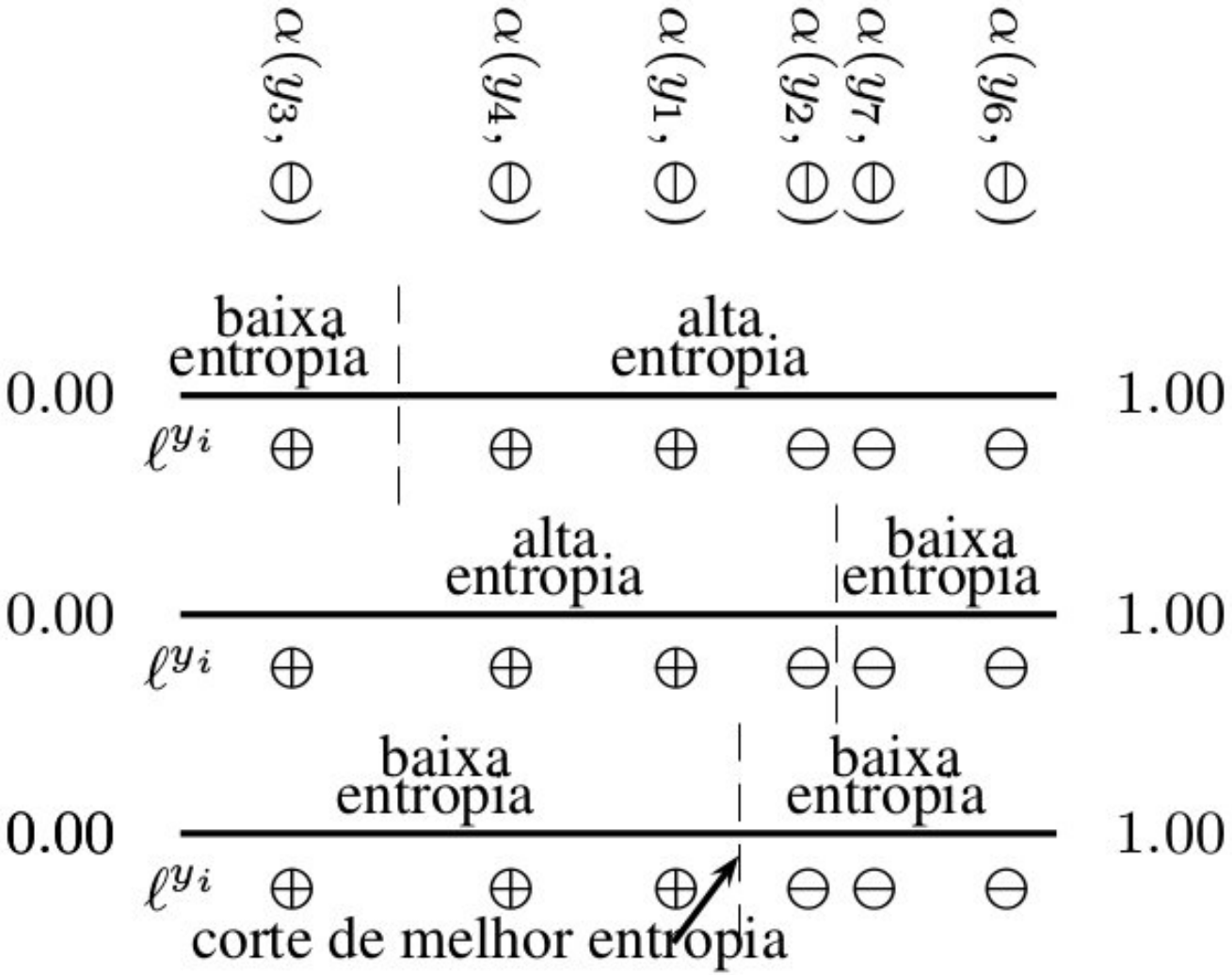
Modelos de Treino: Projeção de Treino

	p					ℓ
	ip4	ip3	ndist	nmatch	fullpath	
y_1	[0.1-0.3]	[0.3-0.5]	[8-10]	[0.2-0.5]	[0.3-0.5]	$\oplus$
y_2	[0.1-0.3]	[0.1-0.3]	[10-14]	[0.1-0.2]	[0.1-0.3]	$\ominus$
y_3	[0.5-0.8]	[0.1-0.3]	[8-10]	[0.1-0.2]	[0.1-0.3]	$\oplus$
y_4	[0.1-0.3]	[0.5-0.8]	[8-10]	[0.2-0.5]	[0.3-0.5]	$\oplus$
y_5	[0.3-0.5]	[0.5-0.8]	[5-8]	[0.5-0.7]	[0.3-0.5]	$\oplus$
y_6	[0.1-0.3]	[0.3-0.5]	[8-10]	[0.5-0.7]	[0.3-0.5]	$\ominus$
y_7	[0.1-0.3]	[0.1-0.3]	[10-14]	[0.1-0.2]	[0.3-0.5]	$\ominus$
y_8	[0.3-0.5]	[0.5-0.8]	[5-8]	[0.5-0.7]	[0.3-0.5]	$\oplus$
x	[0.1-0.3]	[0.1-0.3]	[8-10]	[0.2-0.5]	[0.1-0.3]	?

$\downarrow$ $\downarrow$ $\downarrow$ $\downarrow$ $\downarrow$

	ip4	ip3	ndist	nmatch	fullpath	
y_1	[0.1-0.3]	–	[8-10]	[0.2-0.5]	–	$\oplus$
y_2	[0.1-0.3]	[0.1-0.3]	–	–	[0.1-0.3]	$\ominus$
y_3	–	[0.1-0.3]	[1-2]	–	[0.1-0.3]	$\oplus$
y_4	[0.1-0.3]	–	–	[0.2-0.5]	–	$\oplus$
y_5	–	–	–	–	–	$\oplus$
y_6	[0.1-0.3]	–	–	–	–	$\ominus$
y_7	[0.1-0.3]	[0.1-0.3]	–	–	–	$\ominus$
y_8	–	–	–	–	–	$\oplus$

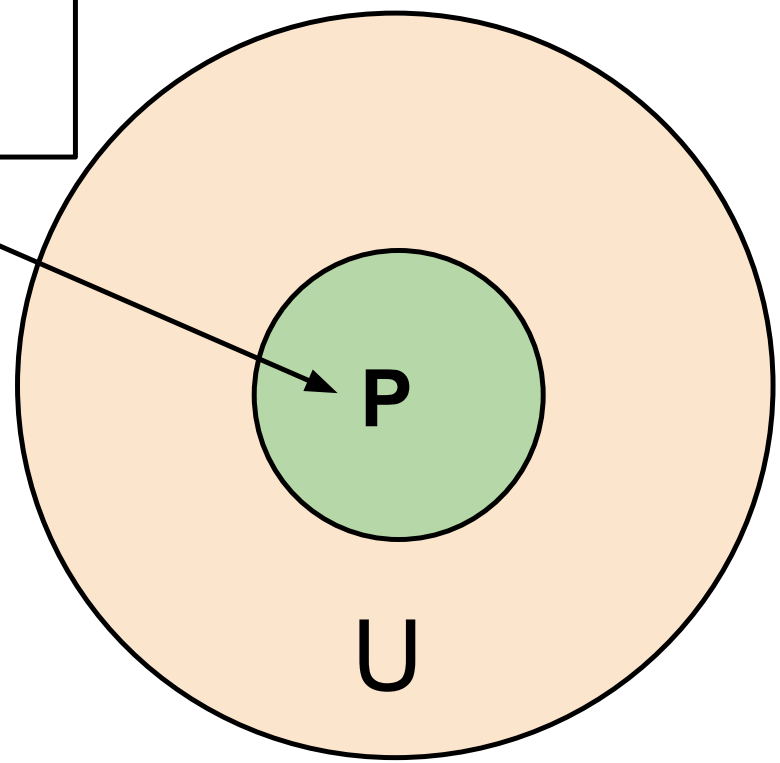
Modelos de Treino: Limiar ótimo



Modelos de Treino: PU

www.exemplo.com.br == exemplo.com.br

exemplo.**gov.br** == exemplo**.br**

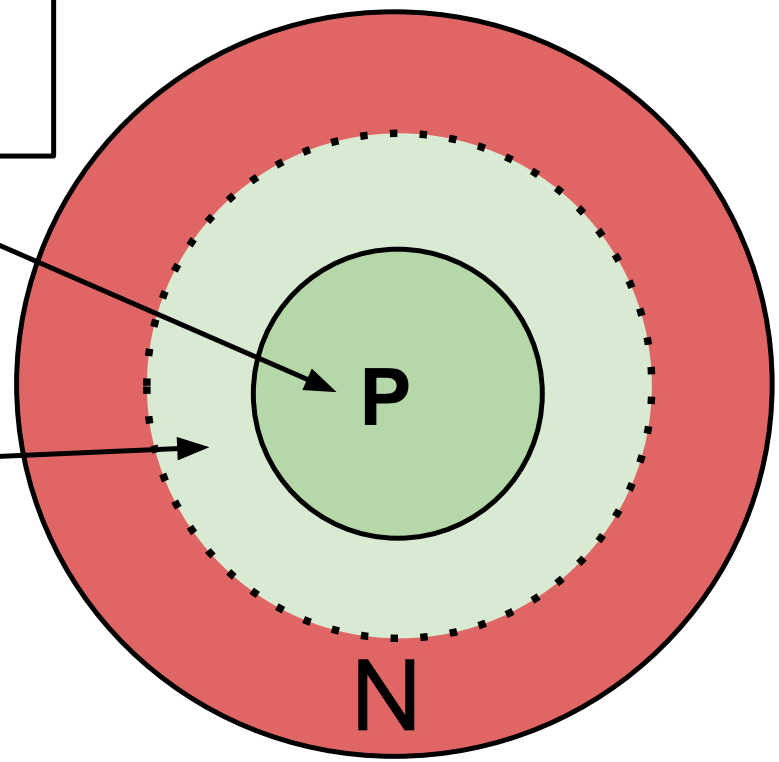


Modelos de Treino: PU -> PN

www.exemplo.com.br == exemplo.com.br

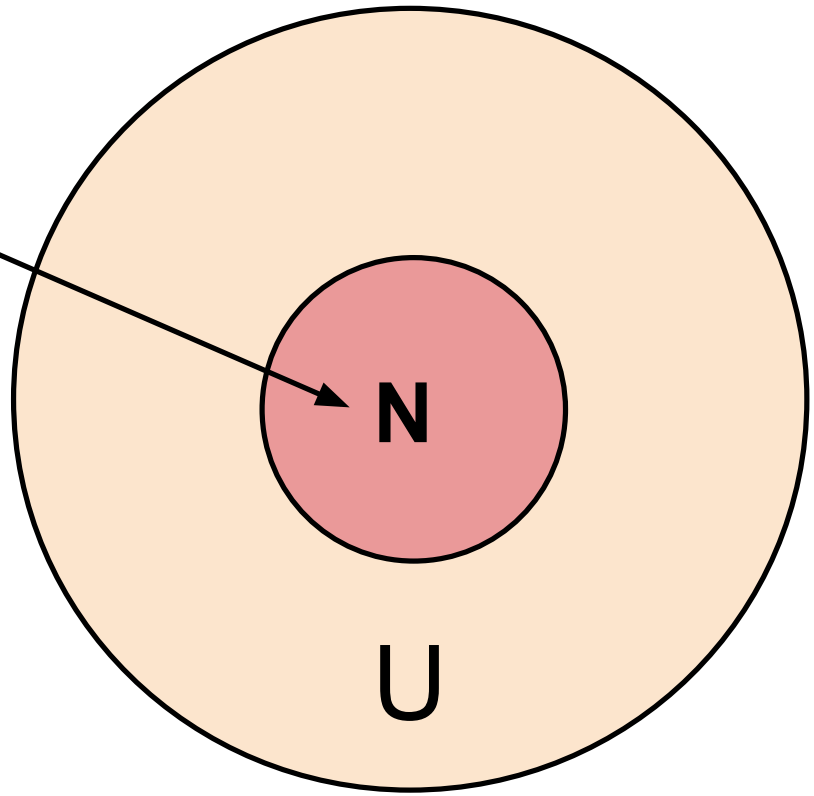
exemplo.**gov.br** == exemplo**.br**

Novos exemplos positivos

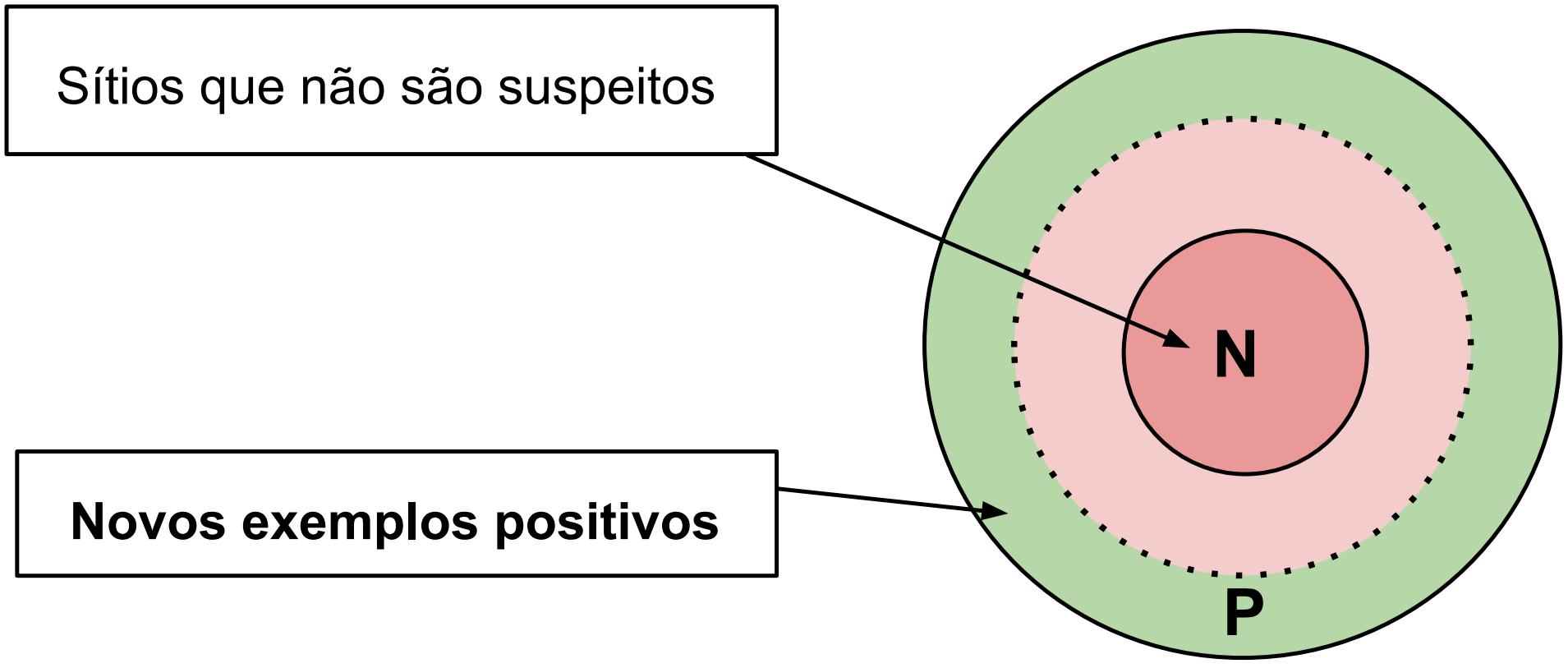


Modelos de Treino: NU

Sítios que não são suspeitos



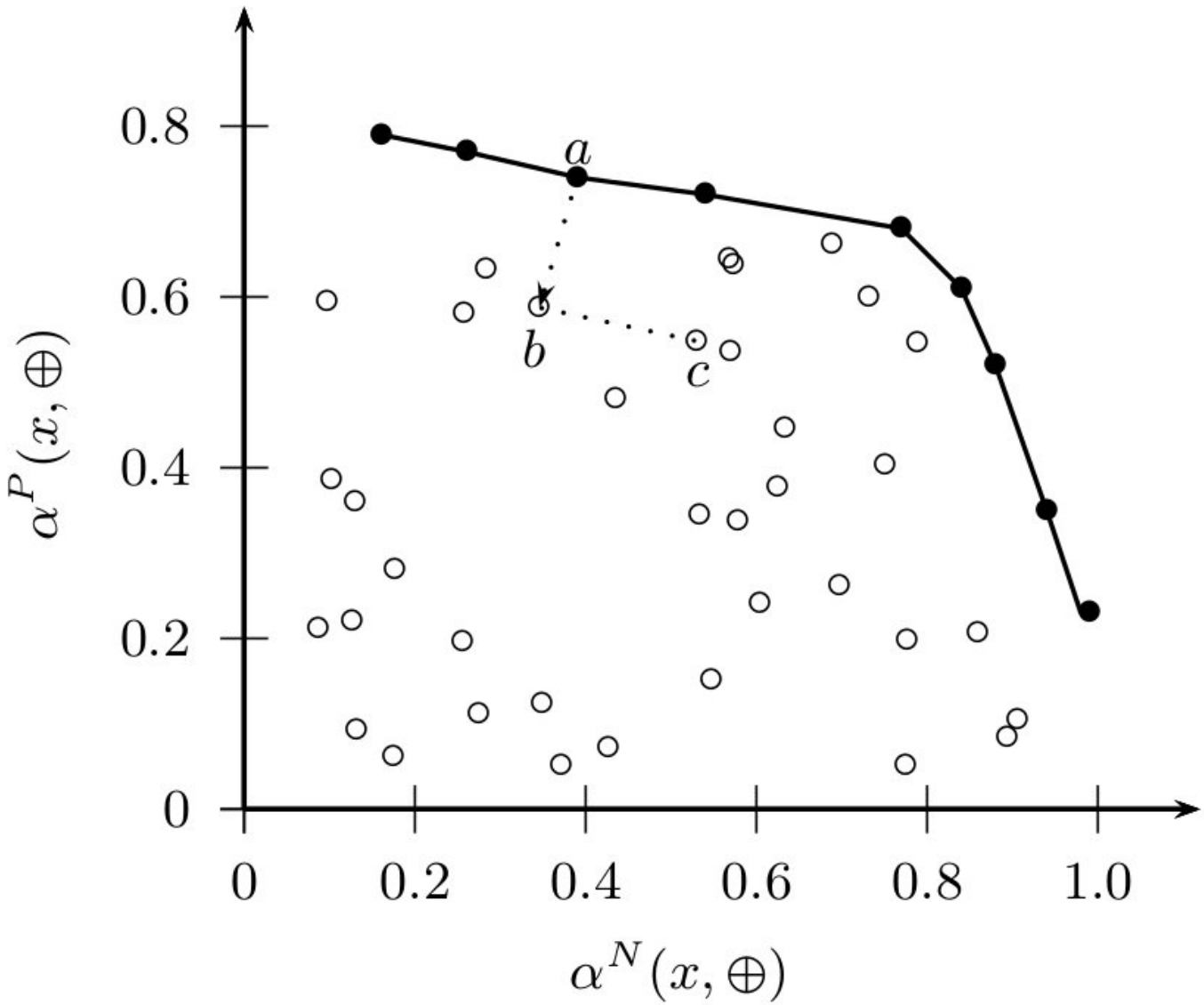
Modelos de Treino: NU -> NP



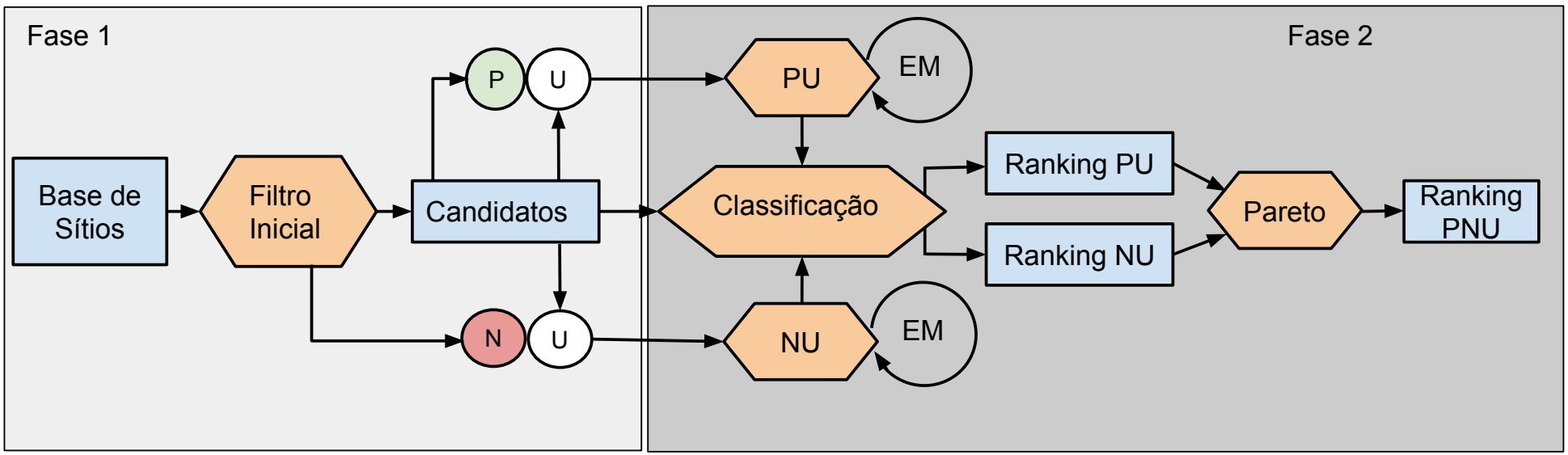
Combinação de Classificadores

- Agregação de resultados (PU + NU)
- Fronteira de Pareto
  - Relação de dominância no espaço de predições
  - Composta por candidatos que se destacam em um classificador ou possuem um balanceamento adequado entre ambos

Combinação de Classificadores



Algoritmo Completo

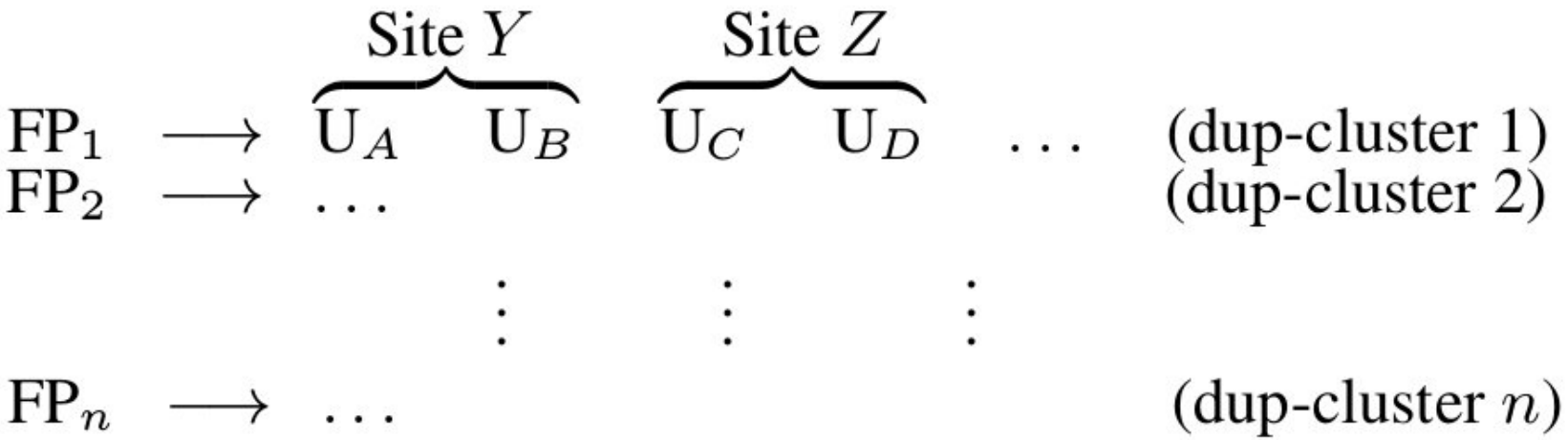


Avaliação Experimental

Coleção

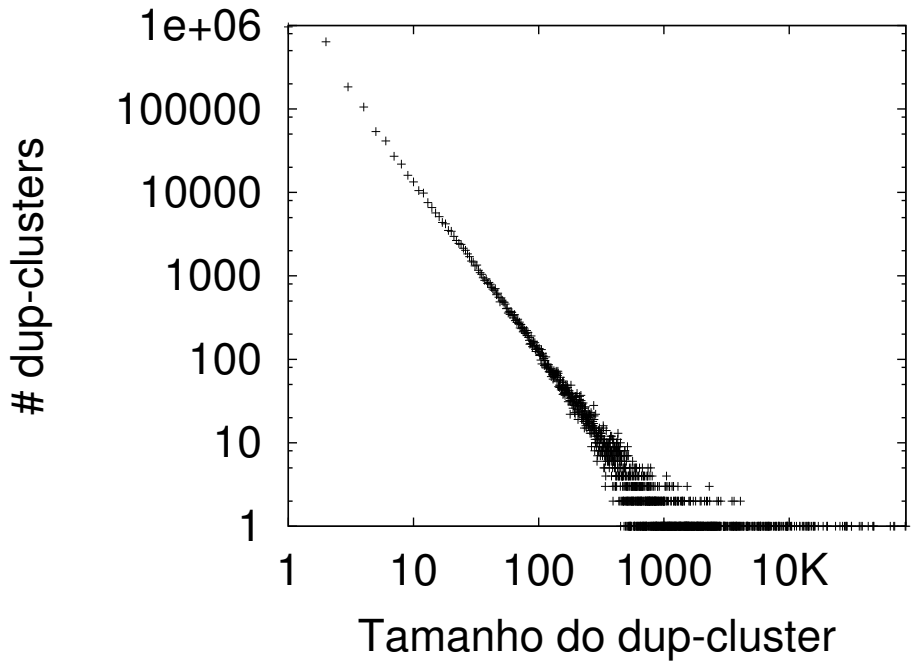
- Coleta entre Setembro e Outubro de 2010
- Usando o coletor InWeb
- Nenhuma restrição ou filtro
- 30 milhões de páginas
- 583,411 sítios web ($\sim 2 \times 10^{11}$ pares possíveis)
- 1.600.000 pares candidatos
- 6.823 pares avaliados como réplicas

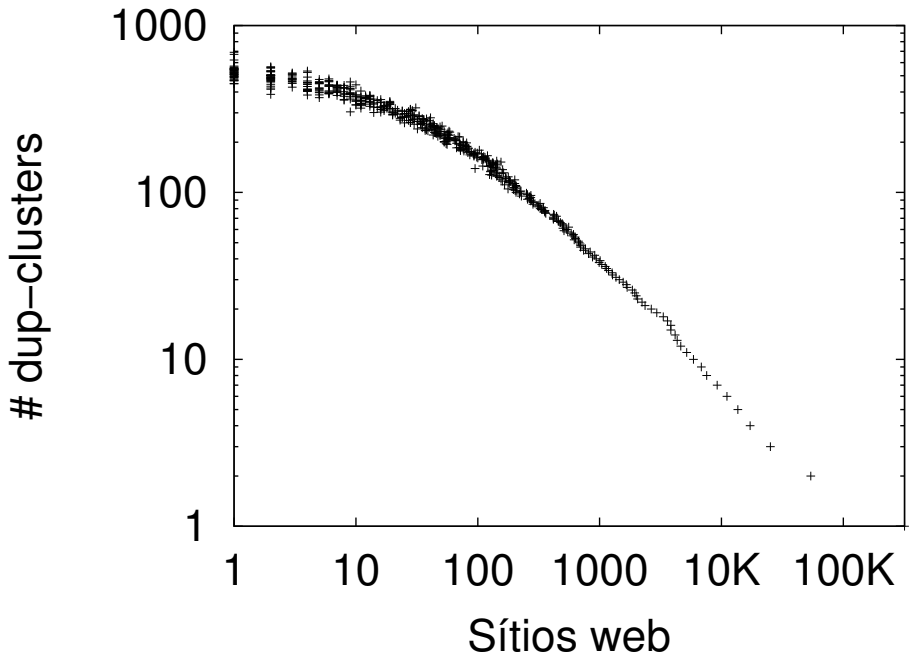
Dup-clusters



- intra-sítios (U_A e U_B)
- inter-sítios (U_A e U_C)

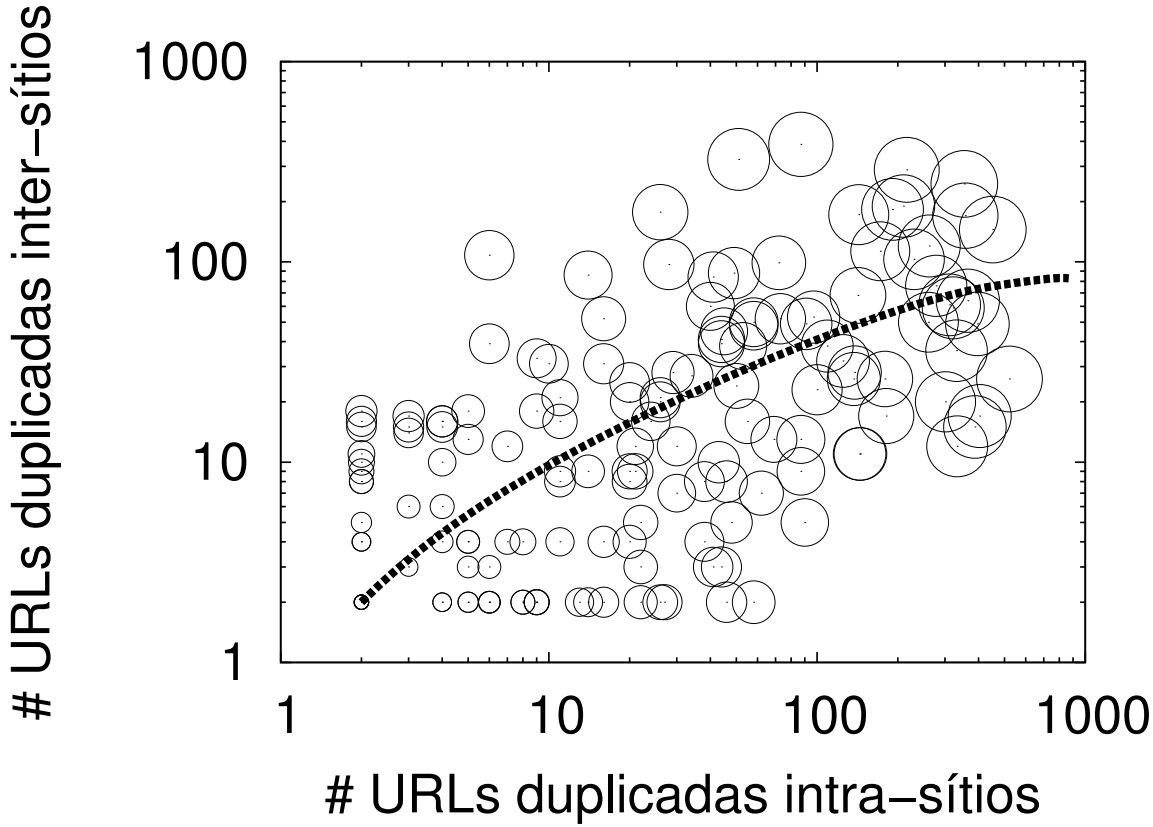
Dup-clusters: Distribuição





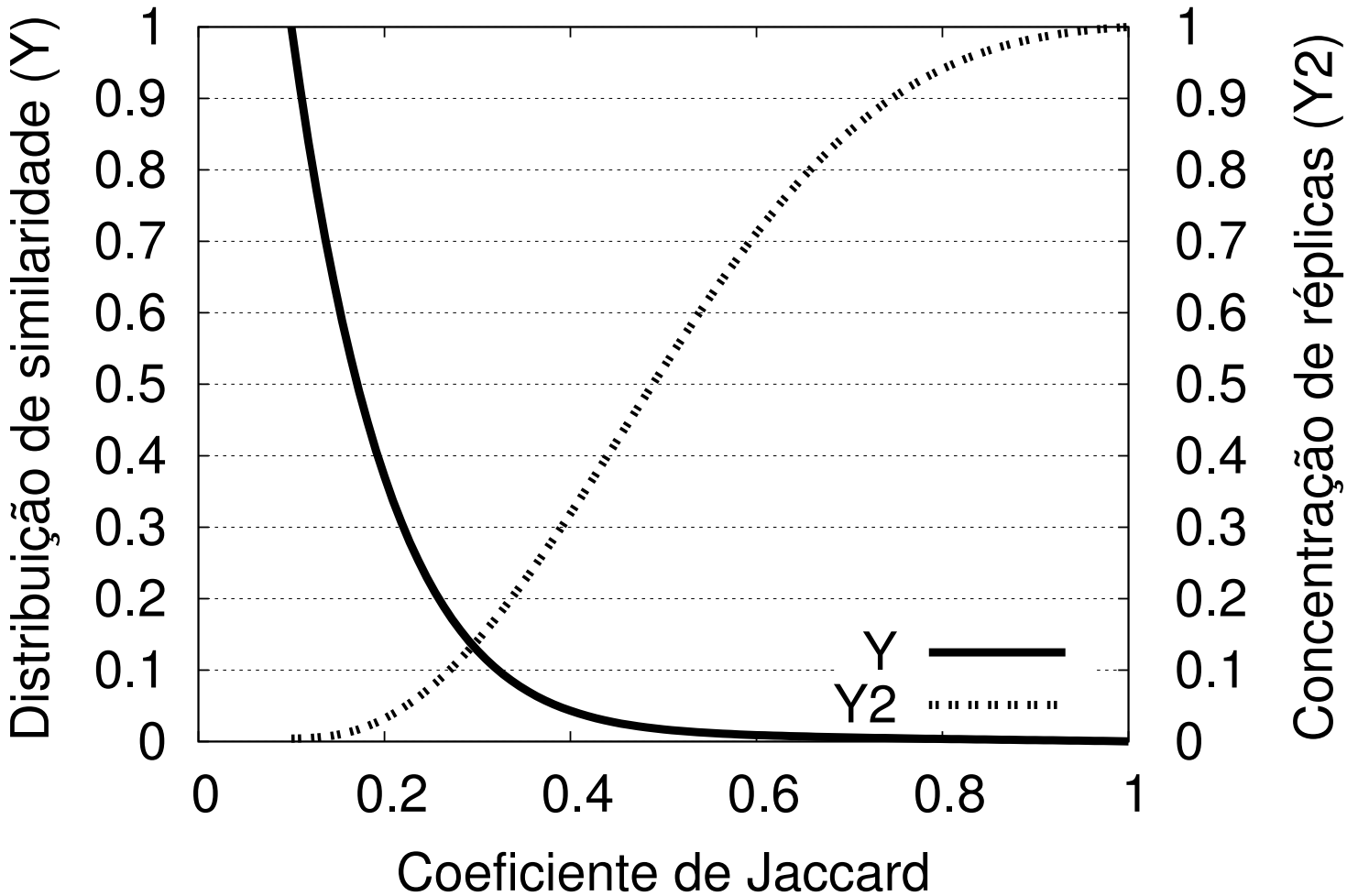
- Muitos dup-clusters contém poucas URLs
- Muitos dup-clusters contém poucos sítios

Inter-sítios vs. Intra-sítios



- Correlação entre o número de URLs duplicadas intra-sítios e inter-sítios por sítio web.

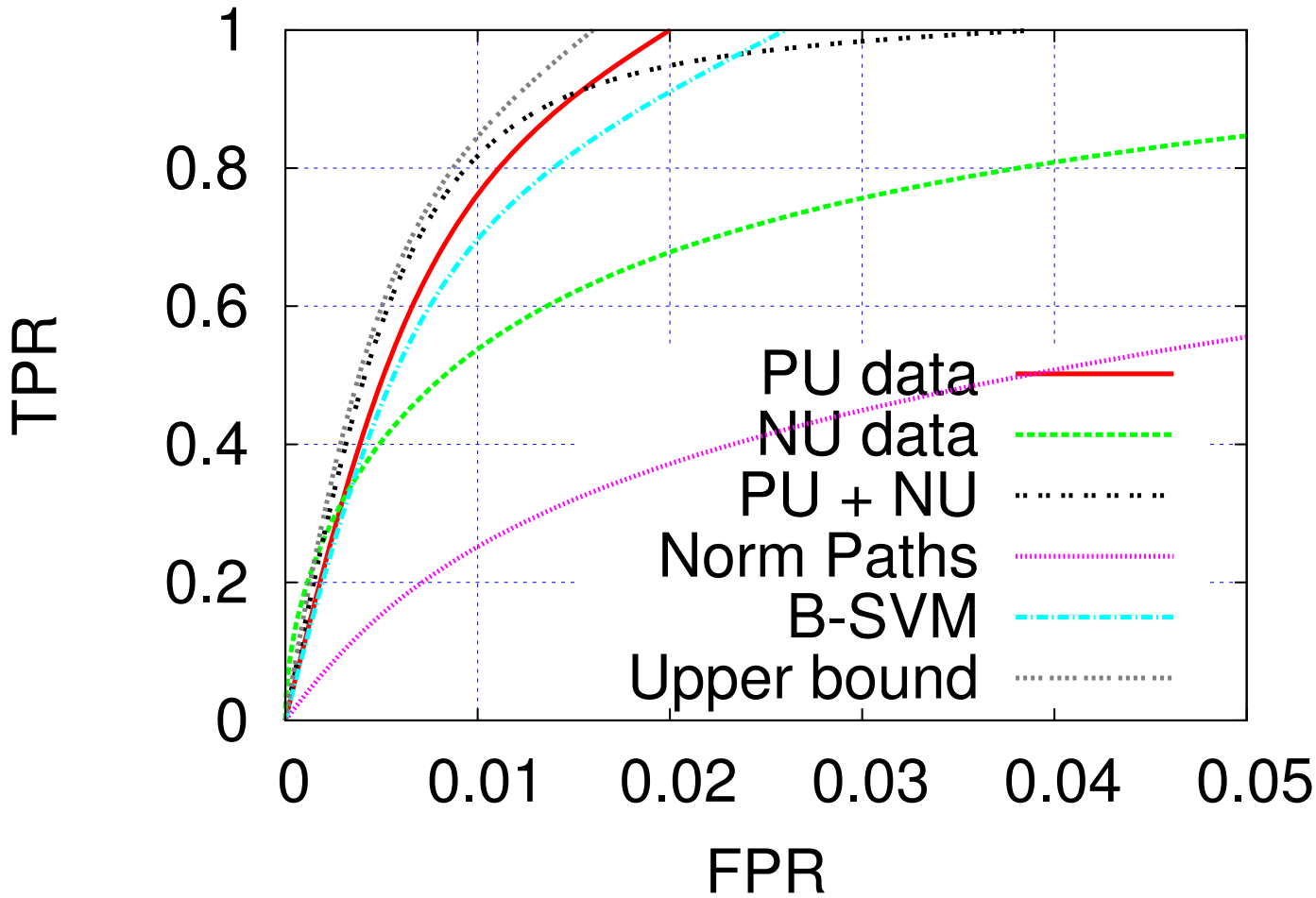
Distribuição de Similaridade



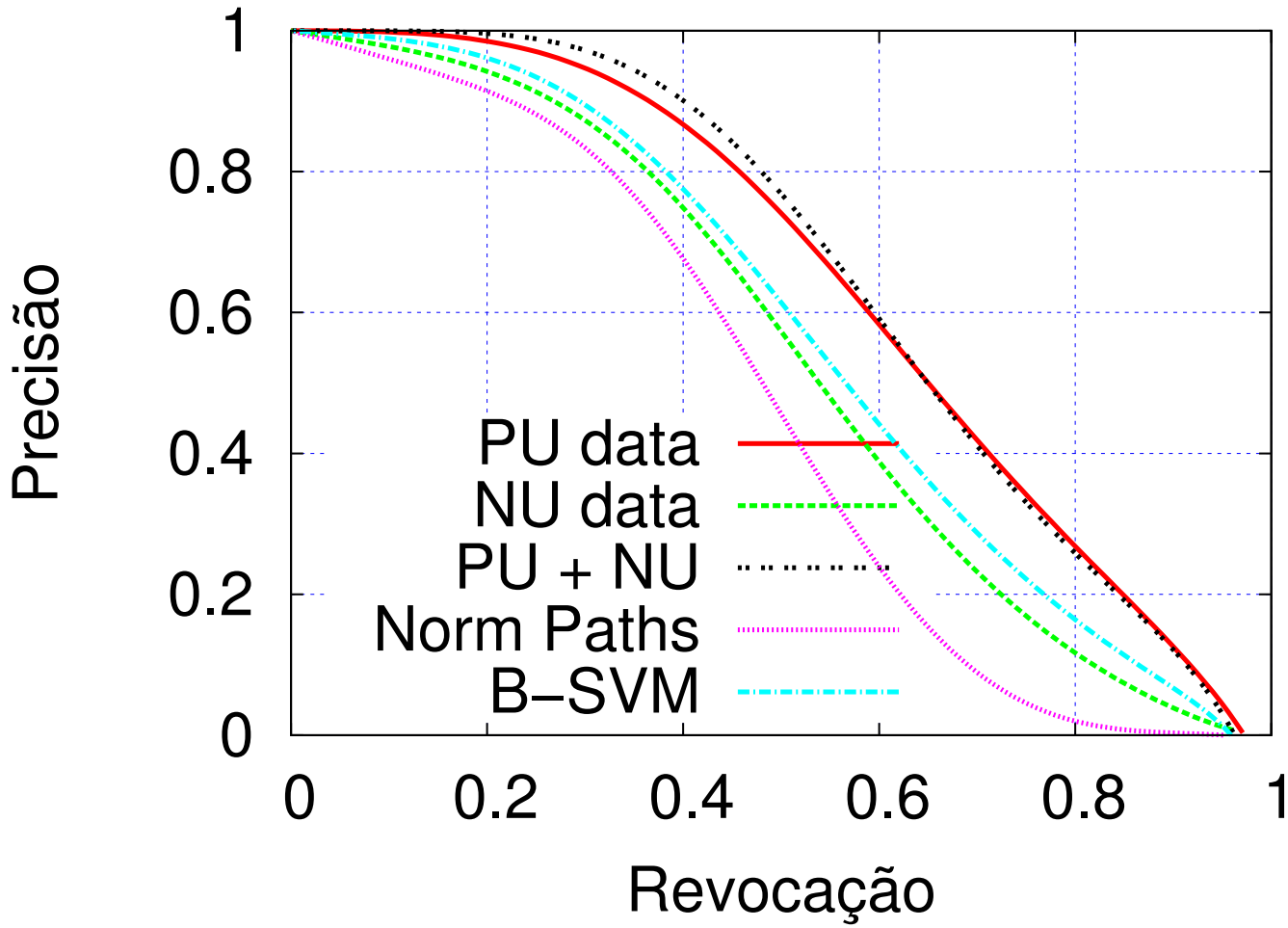
Baselines

- NormPaths
  - Detecção de réplicas de sítios
- B-SVM (*Biased SVM*)[Liu et al.,2003]
  - Estado-da-arte em aprendizado semi-supervisionado com dados PU
- Limite Superior (Upper bound)
  - Classificação com treino rotulado manualmente (Gabarito)

Resultados: ROC



Resultados: Precisão e Revocação



Resultados: Análise de Features

	AUC (Area Under the Curve)			
	Individualmente		Todas exceto	
Features	PU	NU	PU	NU
ip3	0.5132	0.5116	0.9706	0.9411
ip4	0.5288	0.5283	0.9701	0.9431
nmatch	0.6565	0.2312	0.9646	0.9391
ndist	**0.7716**	**0.6528**	0.9631	0.9272
fullpath	**0.7187**	**0.6274**	0.9550	0.9206
$Todas$	0.9908	0.9688	0.9908	0.9688

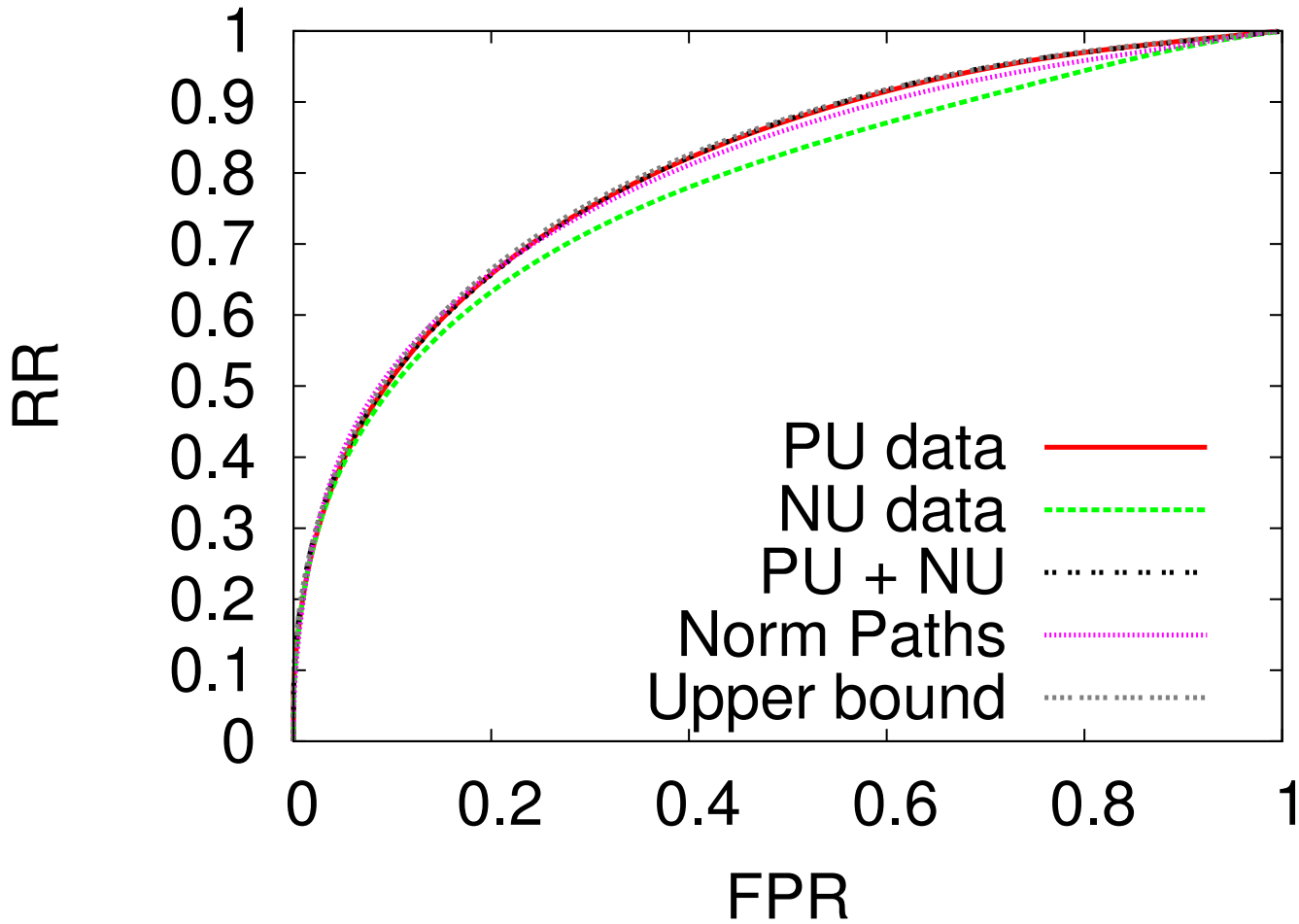
Resultados: Análise de Features

	AUC (Area Under the Curve)			
	Individualmente		Todas exceto	
Features	PU	NU	PU	NU
ip3	**0.5132**	0.5116	0.9706	0.9411
ip4	**0.5288**	0.5283	0.9701	0.9431
nmatch	0.6565	**0.2312**	0.9646	0.9391
ndist	0.7716	0.6528	0.9631	0.9272
fullpath	0.7187	0.6274	0.9550	0.9206
$Todas$	0.9908	0.9688	0.9908	0.9688

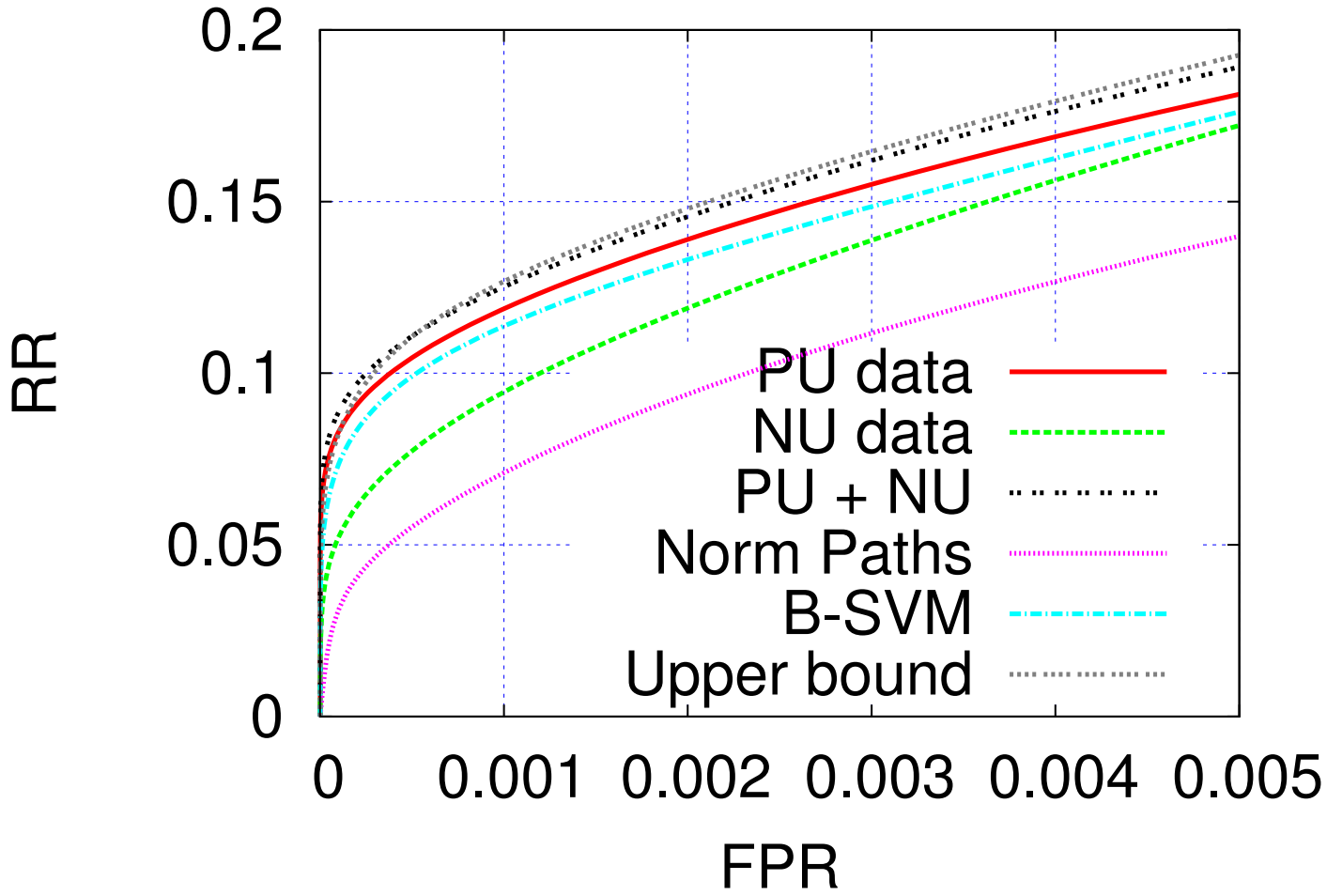
Resultados: Análise de Features

	AUC (Area Under the Curve)			
	Individualmente		Todas exceto	
Features	PU	NU	PU	NU
ip3	0.5132	0.5116	**0.9706**	0.9411
ip4	0.5288	0.5283	**0.9701**	0.9431
nmatch	0.6565	0.2312	**0.9646**	0.9391
ndist	0.7716	0.6528	**0.9631**	0.9272
fullpath	0.7187	0.6274	0.9550	0.9206
$Todas$	0.9908	0.9688	0.9908	0.9688

Resultados: Redução



Resultados: Redução



Resultados: Taxa de Detecção

Algorithms	Número de Candidatos			
	10 + 1	100 + 1	1,000 + 1	10,000 + 1
PU data	0.9900	0.9891	0.9615	0.7656
NU data	0.6590	0.4429	0.4300	0.4287
PU + NU	**1.0000**	**1.0000**	**0.9905**	**0.8272**
NormPaths	**0.6619**	**0.4192**	**0.4105**	**0.4088**
B-SVM	0.9805	0.9622	0.9349	0.7034
Upper bound	1.0000	0.9901	0.9586	0.7555

Resultados: Taxa de Detecção

Algorithms	Número de Candidatos			
	10 + 1	100 + 1	1,000 + 1	10,000 + 1
PU data	0.9900	0.9891	0.9615	0.7656
NU data	**0.6590**	**0.4429**	**0.4300**	**0.4287**
PU + NU	1.0000	1.0000	0.9905	0.8272
NormPaths	**0.6619**	**0.4192**	**0.4105**	**0.4088**
B-SVM	0.9805	0.9622	0.9349	0.7034
Upper bound	1.0000	0.9901	0.9586	0.7555

Resultados: Redução Inter e Intra

		FPR		
		0.000	0.001	0.005
# URLs duplicadas	6,948,501	–	–	–
# URLs intra-sítios	6,514,746	–	–	–
$\rightarrow \ \epsilon = 0.0$	293,374	–	–	–
$\rightarrow \ \epsilon = 0.1$	1,628,685	–	–	–
# URLs inter-sítios	843,526	555,880	865,384	1,331,215
# inter $\cap$ intra	409,771	286,485	446,182	758,927
$\rightarrow \epsilon = 0.0$	64,947	38,835	70,232	110,284
$\rightarrow \epsilon = 0.1$	151,683	98,377	170,827	388,022
RR intra	0.9376	–	–	–
$\rightarrow \epsilon = 0.0$	0.0422	–	–	–
$\rightarrow \epsilon = 0.1$	0.2344	–	–	–
RR inter	0.1213	0.0791	0.1245	0.1916
RR inter + intra	1.0000	–	–	–
$\rightarrow \epsilon = 0.0$	0.1541	0.1166	0.1567	**0.2179**
$\rightarrow \epsilon = 0.1$	0.3340	0.3002	**0.3343**	0.3701

Conclusões

- Foi proposto um novo algoritmo para detecção de réplicas de sítios web em bases de máquinas de busca
- O método semi-supervisionado é capaz de criar de bons modelos de treino
- O algoritmo proposto obteve resultados superiores aos baselines
- A combinação com técnicas de detecção intra-sítios melhora a taxa de redução de duplicatas

Trabalhos Futuros

- Estudo de outras características como a conectividade entre sítios
- Criar estratégias para reavaliação de sítios replicados
- Propor abordagem para escolha justa de qual sítio deve ser eliminado da base (Fraudes).

Contribuições

- Desenvolvimento de novas técnicas de aprendizado de máquina para o problema de detecção de réplicas de sítios
- Coleção com dados a respeito de replicação
  - Sítios coletados, rotulados e modelos de treino
- Artigo a ser submetido
  - WWW 2015 (10 de Novembro)

Contribuições

- Artigo aceito no SPIRE 2013
  - *Learning to Schedule Webpage Updates Using Genetic Programming*
- Artigo em avaliação no *Information Retrieval Journal*
  - *A Genetic Programming Framework to Schedule Webpage Updates*

Obrigado!