Investigando o uso de Características na Detecção de URLs Maliciosas

Gustavo Ferreira da Silva

Câmpus Pato Branco

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS PATO BRANCO

MINISTÉRIO DA EDUCAÇÃO

BRASIL
GOVERNO FEDERAL

URL

- Sigla para Uniform Resource Locator.
- Composta da seguinte forma:
- <esquema>:<parte-especifica-do-esquema>

Composição da URL

- http:// icomp.ufam .edu. br

Esquema

Nome do Domínio Secondary Level Domain (SLD)

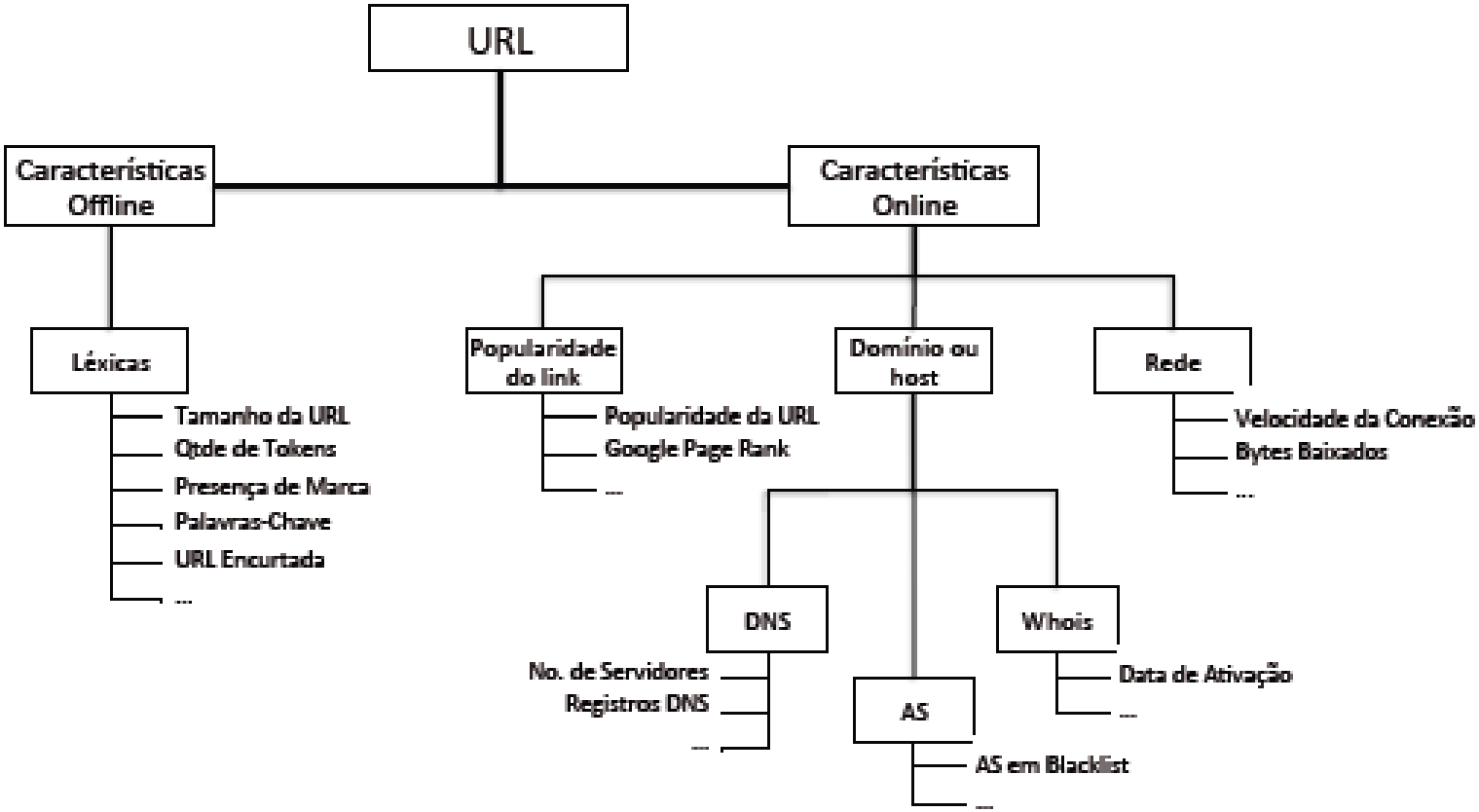
Generic Top Level Domain (GTLD)

Country Code (CCTLD)

Componentes de uma URL

Componente	Exemplo
URL	http://icomp.ufam.edu.br/inst/hstart.php?id=664&logon=141
Nome do domínio	icomp.ufam.edu.br
Caminho	inst/hstart.php
Sub diretório	inst
Nome do arquivo	hstart
Extensão do arquivo	php
Argumento	id=664&logon=141

Agrupamento das URLs



Algoritmos de Aprendizagem de Máquina

- SVM
- Naive Bayes
- Árvore de Decisão (J.48)
- KNN

Características Léxicas

Características Léxicas					
Nome	**Descrição**	**Nome**	**Descrição**	**Nome**	**Descrição**
qt_dom_ponto	Qtde de (.) no domínio	*qt_dom_hifen*	Qtde de (-) no domínio	*qt_dom_underline*	Qtde de (_) no domínio
qt_url_ponto	Qtde de (.) na URL	*qt_url_barra*	Qtde de (/) na URL	*qt_url_interrog*	Qtde de (?) na URL
qt_url_igualdade	Qtde de (=) na URL	*qt_url_hifen*	Qtde de (-) na URL	*qt_url_underline*	Qtde de (_) na URL
qt_url_arroba	Qtde de (@) na URL	*qt_url_ecomerc*	Qtde de (&) na URL	*qt_url_exclam*	Qtde de (!) na URL
qt_url_til	Qtde de (~) na URL	*comp_dominio*	Comprimento do domínio	*comp_url*	Comprimento da URL
qt_dir_ponto	Qtde de (.) no diretório	*qt_dir_barra*	Qtde de (/) no diretório	*qt_dir_interrog*	Qtde de (?) no diretório
qt_dir_igualdade	Qtde de (=) no diretório	*qt_dir_hifen*	Qtde de (-) no diretório	*qt_dir_underline*	Qtde de (_) no diretório
qt_dir_arroba	Qtde de (@) no diretório	*qt_dir_exclam*	Qtde de (!) no diretório	*qt_dir_til*	Qtde de (~) no diretório
qt_arq_ponto	Qtde de (.) no arquivo	*qt_arq_interrog*	Qtde de (?) no arquivo	*qt_arq_igualdade*	Qtde de (=) no arquivo
qt_arq_hifen	Qtde de (-) no arquivo	*qt_arq_underline*	Qtde de (_) no arquivo	*qt_arq_arroba*	Qtde de (@) no arquivo
qt_arq_exclam	Qtde de (!) no arquivo	*qt_arq_til*	Qtde de (~) no arquivo	*qt_par_ponto*	Qtde de (.) no parâmetro
qt_par_barra	Qtde de (/) no parâmetro	*qt_par_interrog*	Qtde de (?) no parâmetro	*qt_par_igualdade*	Qtde de (=) no parâmetro
qt_par_hifen	Qtde de (-) no parâmetro	*qt_par_underline*	Qtde de (_) no parâmetro	*qt_par_arroba*	Qtde de (@) no parâmetro
qt_par_ecomerc	Qtde de (&) no parâmetro	*qt_par_exclam*	Qtde de (!) no parâmetro	*qt_par_til*	Qtde de (~) no parâmetro
qt_params	Qtde de parâmetros na URL	*pres_tld_arg*	Presença de TLD no argumento da URL	*comp_diretorio*	Comprimento do diretório da URL
comp_arquivo	Comprimento do arquivo na URL	*comp_params*	Comprimento dos parâmetros da URL	—	—

Características de DNS

Nome	Descrição
ip_asssociado	No. de IPs resolvidos
sn_associado	No. de servidores de nome resolvidos
data_tempo_ativo	Tempo (em dias) de ativação do domínio

Características Especiais

Nome	Descrição
mal_phi	Presença em listas de Phishing ou Malware
rank_google	Page Rank do Google
presenca_marca	Presença de marca
rank_alexa	Page Rank do Alexa
geo_localizacao	Localização geográfica do domínio
rbl_check	Presença do domínio em RBL (*Real-time Blackhole List*

Ambiente Utilizado

- Windows 7 64 bits, 4 GB de memória RAM, disco de 500 GB e um processador Intel Core i5, 2.3 Ghz
- Intel Core 7 de 3.4 Ghz, com 8 GB de memória RAM, disco de 500 GB e plataforma Linux, distribuição Ubuntu 14.04
- Weka 3.6.10

Bases de Dados Utilizadas

- DMOZ (URLs Benignas)
- PhishTank (URLs Maliciosas)
- Shalla's Blacklist (Agrupamentos de diversas URLs em categorias destinadas a filtros. Também destinada a URLs Maliciosas).

Quantidade de URLs utilizadas

Treinamento e Ajuste

- 20.092 – Total
- 10.046 – Base DMOZ
- 10.046 – Base PhishTank

Quantidade de URLs utilizadas

Teste

- 20.000 – Base DMOZ e PhishTank de forma igualitária
- 20.000 – Base DMOZ e Shella’s forma igualitária

Medidas de Desempenho

- Taxa de Detecção: VP / (VP+FN)
- Taxa de Precisão: (VP+VN) / (VP+VN+FP+FN)
- Taxa de Falso Alarme: FP / (FP+VN)

- VN (Verdadeiro Negativo)
- VP (Verdadeiro Positivo)
- FP (Falso Positivo)
- FN (Falso Negativo)

Métrica empregada

- 10 partições para os 4 classificadores.

Comparação

Classificador	Naive Bayes	SVM	Árvore de Decisão	KNN
Parâmetros Ajustados	-	*C=50, Kernel* Polinomial, Grau do Polinômio= 1.0	Fator de Confiança= 0,25	KNN=1
Taxa de Precisão	76,00%	91,40%	95,10%	94,90%
Taxa de Detecção	66,35%	91,43%	95,11%	94,91%
Falso Alarme	33,60%	8,60%	4,90%	5,10%

Hipóteses

- **H1. Existe alguma influência do formato da URL na extração das características e, consequentemente, no processo de avaliação?**
- **H2. Todas as características extraídas são realmente necessárias no processo de detecção de URLs?**
- **H3. Grupos de características permitem resultados adequados, e até melhores, no processo de detecção de URLs se comparados com características individuais.**
- **H4. A importância das características depende da base onde as URLs são coletadas.**

Provas H1

Características	Shalla's (Dez. 2014)	PhishTank (Nov. 2013)
comp_url	31,0367	54,1183
comp_dominio	14,3585	19,0749
comp_diretorio	6,1998	21,6629
comp_arquivo	5,3814	7,1557
comp_params	3,097	6,2263
qt_tok_dir_barra	1,7573	2,9908
sn_assoc	2,4993	1,8551
rank_alexa	129,06	16430,6377
rank_google	2,0873	0,359

Provas H3

Conjunto	Descrição	Conjunto	Descrição
A	Composto pelas 3 características baseadas em informações obtidas do DNS	B	Composto pelas 6 características denominadas especiais
C	Composto pelas 15 características léxicas mais comuns [1]	D	Composto por 32 características léxicas variáveis (não aparecem em todas as bases de dados)
E	Composto por todas as características léxicas	A+B+C	Composto pelos grupos A, B e C, totalizando 24 características
A+B+D	Composto pelos grupos A, B e D, totalizando 41 características	Todos	Todas as 56 características

Provas H2 e H4

DMOZ/PhishTank		DMOZ/Blacklist	
Característica	**InfoGain**	**Característica**	**InfoGain**
rank_google	0.390339	*qt_tok_url_barra*	0.70944
data_tempo_ativo	0.348147	*data_tempo_ativo*	0.452587
geo_localizacao	0.228261	*qt_tok_dir_barra*	0.276703
qt_tok_dom_ponto	0.158864	*geo_localizacao*	0.2612
qt_tok_url_barra	0.14602	*qt_tok_dom_ponto*	0.237529
qt_tok_dir_barra	0.118791	*comp_arquivo*	0.233644
comp_diretorio	0.115648	*rank_alexa*	0.180241
comp_url	0.105754	*comp_dominio*	0.176853
qt_tok_url_ponto	0.096124	*ip_assoc*	0.162585
rbl_check	0.082557	*qt_tok_url_ponto*	0.133253
comp_arquivo	0.074864	*rank_google*	0.120826
comp_dominio	0.073538	*presenca_marca*	0.102384
sn_assoc	0.068434	*sn_assoc*	0.067489
presenca_marca	0.056488	*rbl_check*	0.067109
rank_alexa	0.051651	*comp_diretorio*	0.06236
ip_assoc	0.038537	*comp_url*	0.052208
qt_tok_dir_ponto	0.035866	*comp_params*	0.037848
comp_params	0.029835	*qt_tok_arq_ponto*	0.032119
qt_tok_url_hifen	0.029591	*qt_tok_dir_hifen*	0.03101
qt_tok_dir_hifen	0.028234	*qt_tok_url_igualdade*	0.02972
qt_tok_url_ecomerc	0.024787	*qt_tok_par_igualdade*	0.029292
qt_tok_dom_hifen	0.02455	*qt_tok_arq_til*	0.022689
qt_tok_par_ecomerc	0.022983	*qt_tok_url_hifen*	0.018402
qt_params	0.022983	*qt_tok_url_interrog*	0.014041