Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

Pablo Augusto da Paz Elleres

Detecção de *Canvas Fingerprinting* em Páginas Web baseada em Modelo Vetorial

Manaus
Março de 2017

Pablo Augusto da Paz Elleres

Detecção de *Canvas Fingerprinting* em Páginas Web baseada em Modelo Vetorial

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito para obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Eduardo Luzeiro Feitosa

Manaus
Março de 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Elleres, Pablo Augusto da Paz
E45d Detecção de Canvas Fingerprinting em Páginas Web baseada em Modelo Vetorial / Pablo Augusto da Paz Elleres. 2017
125 f.: il. color; 31 cm.

Orientador: Eduardo Luzeiro Feitosa
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Website Fingerprinting. 2. Canvas Fingerprinting. 3. Recuperação da Informação. 4. Método Vetorial. 5. Similaridade. I. Feitosa, Eduardo Luzeiro II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Detecção de Canvas Fingerprinting em Páginas Web Baseada no Modelo Vetorial"

PABLO AUGUSTO DA PAZ ELLERES

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Luzeiro Feitosa - PRESIDENTE

Prof. Eduardo James Pereira Souto - MEMBRO INTERNO

Prof. Rafael Roque Aschoff - MEMBRO EXTERNO

Manaus, 31 de Março de 2017

Dedicatória

Dedico às pessoas mais importantes da minha vida e as quais sem seus apoios de modo incondicional jamais conseguiria ter êxito, minha família!

"Devemos gritar em silêncio que os melhores dias estão por vir, enfrentar os períodos mais tristes da vida não como pontos finais, mas como vírgulas para continuarmos a escrever nossa trajetória".

Augusto Cury (O Vendedor de Sonhos)

Agradecimentos

Agradeço a Deus, por ter me iluminado e acompanhado nessa caminhada, dando-me forças e discernimento para seguir nos momentos de dificuldade.

Agradeço de modo especial a minha família: Mário e Angela Elleres (meus pais), a minha avó Noêmia (*in memoriam*) que ensinou-me a ler e escrever, a tia Socorro, a meu irmão Diego (*in memoriam*) que incentivou-me na busca da realização de meus sonhos. Agradeço a vocês por acreditarem em minha capacidade, graças a vocês tive forças para continuar e realizar mais este sonho. Eu os amo e vocês são tudo em minha vida!

Agradeço ao profissional mais dedicado, simples e competente, que sempre teve toda paciência em conduzir-me durante esta jornada. E certamente é o responsável direto por esta conquista, meu orientador, Prof. Eduardo Luzeiro Feitosa. Lhe serei eternamente grato, não somente pelo aprendizado da pesquisa, mas também pelo aprendizado de vida, por ensinar-me a ser mais: humano, simples e respeitoso para com os que nos cercam.

Um agradecimento especial ao professor e colega, Ednaldo Coelho Pereira, que incentivou e me inspirou como profissional competente e dedicado que é.

Agradeço também a todos os professores do PPGI pelos conhecimentos transmitidos e pelas amizades aqui conquistadas.

A todos os meus colegas de mestrado e, de forma especial, aos amigos que fiz nesta jornada: Adria, Adriana, Caio Gregoratto, Carlos, Jordan, Helen Sobrinho, Maria Azevedo, Michel, Rayol Neto e Thais, pelo companheirismo que sempre tiveram para comigo nos momentos de maior dificuldade e principalmente pela atenção ao compartilharem seus conhecimentos.

Aos amigos Lene, Blaíse, Aladilson, Wanderson, Wallace e Orley que mesmo distantes sempre contribuíram, de maneira direta dando palavras de apoio e de garra durante essa jornada, descrevo aqui os meus sinceros agradecimentos e reconhecimento.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo auxílio financeiro. A todos vocês que me ajudaram, meus sinceros agradecimentos!

Resumo

Fingerprinting é a técnica aplicada com vistas a identificar ou reidentificar um usuário/dispositivo por intermédio de um conjunto de atributos como: o tamanho da tela do dispositivo, a identificação do endereço IP, as versões dos softwares instalados, assim como por meio de outras características existentes no processo de comunicação da Web. A técnica é conhecida pela nomenclatura de *Website fingerprinting* e tem sido utilizada como mecanismo de marketing/vendas de produtos, mas pode muito bem ser empregada como medida de segurança na autenticação de usuários. A questão é que ela pode e deve ser considerada uma ameaça potencial a privacidade dos usuários na Web, já que dados pessoais e sigilosos podem ser capturados e empregados para fins maliciosos. Atualmente uma técnica que utiliza renderização de imagens, denominada Canvas *fingerprinting*, também tem sido utilizada para burlar a privacidade dos usuários de websites. Este trabalho apresenta um método que emprega técnicas de recuperação da informação (via método vetorial), para realizar a detecção de scripts Canvas *Fingerpriting* em páginas Web. O método consiste em realizar o cálculo da similaridade entre uma base com 100 consultas reconhecidamente ligadas à Canvas *Fingerpriting* e bases de dados com páginas tidas como benignas e malignas. O resultado encontrado mostrou que níveis altos de similaridades com uma base de Canvas (97%), uma base de páginas *phishing* (87%) e uma base com páginas do diretório DMOZ (87%).

Palavras-chave: *Website Fingerprinting*, Canvas *Fingerprinting*, Recuperação da Informação, Método Vetorial, Similaridade.

Abstract

Fingerprinting is a technique applied in order to identify or re-identify a User/ device via a set of attributes such as the size of the device's screen, IP address identification, the versions of the software installed as well as through other existing features in the process Web communication. The technique is known in Nomenclature website fingerprinting and it has been used as a mechanism for marketing/product sales, however, its development aims to serve as a measure security of user authentication. The question is As it is considered a potencial threat to Web privacy, since personal and sensitive data can be captured and used for malicious purposes in various types of attacks and fraud. The point is that it may and should be considered a potential threat to the privacy of users on the Web, since personal and sensitive data can be captured and used for malicious purposes. Currently a technique that uses image rendering, called Canvas fingerprinting, has also been used for the same purposes as the previous one. This work presents a method that uses information retrieval techniques (via vectorial method) to perform the detection of Canvas Fingerprinting scripts in Web pages. The method consists in calculating the similarity between a base with 100 queries from a Canvas Fingerprinting database and a set of web pages labeled as benign and malignant. The result found showed high levels of similarities with a canvas base (97 %), a base of phishing pages (87 %) and a base with DMOZ directory pages (87 %).

Keywords: Website Fingerprinting, Canvas Fingerprinting, Information Retrieval,Vectorial Method, Similarity.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Por proporcionar acesso quase sem restrições a uma variada gama de serviços e informações, a Web se tornou o ambiente multimídia mais utilizado em todo o mundo, influenciando não só as relações pessoais como também comerciais nas últimas duas décadas. Entretanto, graças a realização de transações de comércio eletrônico e as atividades de prestação de serviços, a Web tornou-se um dos alvos preferidos de ataques à segurança e a privacidade dos usuários. Em geral, o objetivo dos vários tipos de ataques e códigos de exploração executados na Web, tendo os navegadores como porta de entrada, é capturar informações privadas de pessoas e empresas como o número do cartão de crédito, conta no banco, logins e senhas, entre outros.

Mas como os usuários podem ser identificados na Web, e ter seus dados coletados, se uma das principais características da Internet é o anonimato? A forma mais popularizada para identificação de usuários na Web é através dos Cookies[1], criados com a finalidade de personalizar a navegação dos usuários, mas que passaram a ser utilizados para registrar informações com fins comerciais. Empresas de publicidade e propaganda on-line, além de atuar diretamente nos anúncios de produtos e serviço na Web, tornaram-se especializadas em rastrear usuários, lucrando fortunas com esta atividade. Em linhas gerais, tais empresas fecham parcerias com vários sites Web e outras empresas com a finalidade de coletar dados dos usuários durante a navegação e, assim, construir perfis detalhados dos

[1] Cookies são pequenos pedaços de texto (arquivos) que os sites fazem o navegador do usuário salvar.

interesses e atividades destes usuários. Este perfil, criado normalmente sem o consentimento dos usuários, gera uma série de preocupações, pois deixa os usuários não só vulneráveis a furtos e fraudes, mas também expõe sua privacidade on-line.

Embora diversos mecanismos de segurança tenham sido desenvolvidos visando acabar ou dificultar tais atividades na Web, incluindo leis anti-Cookie na Europa e Estados Unidos (The Wall Street Journal)[1], ainda existem outras formas de obter dados dos usuários durante a navegação. Atualmente, técnicas de *Website Fingerprinting* têm sido exploradas com a finalidade de identificar/reidentificar usuários, de maneira única, por meio dos dispositivos utilizados por eles. Endereço IP, tamanho da tela do dispositivo, versões de softwares instalados, tipos de fontes e uma diversidade de outras características observáveis durante processo de comunicação são utilizadas para isso.

1.1 Motivação

O simples ato de navegar e acessar sites e serviços de forma anônima, ou através do uso de pseudônimos, já não garante mais privacidade. Isso porque, querendo ou não, tudo o que o usuário acessa fica registrado, seja através do seu endereço IP e cabeçalhos do protocolo HTTP ou através dos Cookies existentes nos sites.

Mesmo que alguns usuários não vejam problemas nestes fatos, a posse de dados confidenciais ou não dos usuários, coletados sem consentimento, por empresas de publicidade on-line ou mesmo por qualquer empresa ou agência governamental, traz consequências potencialmente desastrosas para a privacidade das pessoas e, em casos mais graves, pode envolvê-las em furtos, fraudes e ataques maliciosos.

A primeira solução para este problema surgiu em 1997, quando a RFC 2109 [2] especificou um padrão para o uso de Cookies, o qual proibia o seu uso por terceiros. Naquela época, nenhum dos dois navegadores existentes, Internet Explorer e Netscape Navigator, adotou a especificação e a proposta nunca foi bem sucedida, apesar de vigorar até hoje.

Em 2005, os desenvolvedores de navegadores começaram a adicionar em seus produtos o modo de navegação privada, a fim de permitir que os usuários optassem por visitar sites sem deixar Cookies de longo prazo. Partindo dessa premissa,

outros desenvolvedores começaram a produzir extensões para os navegadores, com vistas a preservar a privacidade dos usuários. *AdBlockPlus*[2], *Ghostery* [3] e *Lightbeam* [4] são bons exemplos.

Porém, mesmo com estes tipos de soluções, as empresas de anúncios e publicidade não ficaram paradas. Agora, aproveitando-se do grande foco dado aos dispositivos móveis, elas modificaram a maneira de rastrear usuários na Web. Fazendo uso de ferramentas de rastreio e reconhecimento capazes de detectar o conjunto hardware e software do dispositivo utilizado pelo usuário, estas empresas de anúncios e publicidade criaram o conceito de *Website Fingerprinting*. De acordo com Flood e Karlsson [3], *Website Fingerprinting* é um conjunto de propriedades de um dispositivo e seu software que são recolhidas a partir de um navegador. Ou seja, as propriedades de hardware e as informações relacionadas ao software identificam determinado dispositivo.

Atualmente, os *Website Fingerprinting* mais comuns tem utilizado tecnologias como JavaScript e ActionScript (Flash). Contudo, existem outras possibilidades capazes de identificador um usuário/dispositivo. É o caso da tecnologia HTML Canvas, que faz uso de propriedades e métodos gráficos. A ideia do Canvas *fingerpriting* é instruir o navegador a desenhar linhas, textos, figuras geométricas, entre outros, que a posteriori são convertidos em um identificador único. Conforme explicitam Ximenes et al. [4], basicamente a estratégia é desenhar uma imagem (com textos e cenas WebGL) usando comandos gráficos do HTML5 através da *tag* <canvas> para, posteriormente, fazer a captura da imagem desenhada. A imagem capturada é usada para construir uma assinatura exclusiva para cada navegador. Essa assinatura é, em seguida, remetida ao servidor, que realiza o rastreamento.

Embora não existam leis que proibam o uso e a execução desse tipo de técnica (mesmo sabendo que elas têm sido comumente executadas pelos mais diversos sites Web), a questão é que os usuários nem sequer tem conhecimento de sua existência ou ignoram o fato de que estão sendo rastreados. Este fato deixa o seguinte questionamento: será que há uma maneira de quantificar o quanto os

[2] https://adblockplus.org
[3] https://ghostery.com
[4] https://addons.mozilla.org/pt-br/firefox/addon/lightbeam

usuários da Web estão vulneráveis aos ataques de *Website Fingerprinting* por meios das propriedades de Canvas, como forma de alertá-los sobre esse tipo de situação?

1.2 Objetivo

O objetivo desta pesquisa é propor um método capaz de detectar e avaliar scripts Canvas *fingerprinting* em páginas Web com base na extração de características (propriedades e métodos) relevantes do conteúdo estático do documento Web, por meio de técnicas de recuperação da informação, a fim de apresentar um ranqueamento destes scripts por níveis de similaridade. Desta forma, é possível conhecer os scripts originais e seu grau de periculosidade para os usuários ao acessarem páginas Web com este tipo de mecanismo.

O método proposto permite não somente uma exploração mais detalhada do problema, mas também uma análise de um conjunto de características relevantes que podem ser extraídas de páginas Web, de acordo com sua capacidade de gerar ataques ou ameaças à privacidade dos usuários web, permitindo sua classificação por níveis de similaridade. Esta classificação por nível de similaridade é interessante não só por oportunizar uma comparação entre os scripts, mas principalmente porque poderá servir para uma futura implementação de um plug-in. Para validar os resultados, uma prova de conceito fora implementada e fez uso de *scripts* obtidos de bases reais da Internet, como, por exemplo, das bases da Universidade de Princenton (base Canvas *fingerprinting*), do Phishtank (base maliciosa de *phishing*), Dmoz (base benigna) e do Alexa.com (sites mais visitados da web).

1.3 Contribuições

A partir dos objetivos definidos foram alcançadas as seguintes contribuições:

- Definição e análise de um conjunto de características capazes de discriminar Canvas *Fingerprinting* em páginas Web que possam ser adicionadas em

políticas, regras, filtros e sistemas de detecção ou, ainda, serem aplicadas como solução complementar a outras técnicas atualmente usadas;

- Criação de um método baseado no cálculo de similaridade do modelo vetorial para detectar Canvas *fingerprinting* em páginas Web.

1.4 Estrutura do Documento

O restante desta dissertação está organizado como segue: O Capítulo 2 apresenta os conceitos concernentes ao tema *Website Fingerprinting*, incluindo definições e classificação, bem como informações sobre a recuperação da informação e os tipos de métodos por ela utilizados. Os trabalhos relacionados são discutidos no Capítulo 3, objetivando apresentar trabalhos acadêmicos e soluções existentes que enfatizam a detecção de *Website Fingerprinting*, bem como trabalhos que tratem não somente da classificação de *scripts fingerprinting*, mas também do uso da tecnologia Canvas para realização desse tipo de ataque. O Capítulo 4 apresenta a proposta da dissertação, na qual destaca-se, de maneira detalhada, uma visão geral do método proposto e a descrição de seus componentes. No Capítulo 5 são descritos todos os experimentos realizados nesta pesquisa, bem como a validação dos resultados obtidos. No capítulo 6 é realizado o fechamento da dissertação, incluindo os trabalhos futuros que podem ser utilizados para a continuação da pesquisa.

Capítulo 2

Conceitos Básicos

Este capítulo apresenta os principais conceitos, definições e classificações sobre *Website Fingerprinting*. Além disso, trata de recuperação da informação e outros conceitos que são discutidos e exemplificados, com o intuito de preparar o leitor para os restante desta dissertação.

2.1 *Website Fingerprinting*

O termo *fingerprinting* ganhou força na área da computação nos anos de 1990 com o surgimento de várias ferramentas especializadas em realizar ataques a redes de computadores. A ideia era de que para efetuar um ataque bem sucedido era necessário descobrir/identificar corretamente a máquina (computador/servidor) alvo, o sistema operacional que ela executava e os aplicativos ativos.

No âmbito Web e foco deste Capítulo, o termo *fingerprinting* veio a tona em 2009, quando Mayer [5] observou que as características de um navegador e seus plugins podiam ser identificadas e o usuário rastreado. Em 2010, Eckersley [6] mostrou que informações (atributos) fornecidas pelo navegador dos usuários eram suficientes para identificar a grande maioria das máquinas que navegam na Internet. Eckersley desenvolveu um algoritmo para investigar o grau em que os navegadores modernos estão sujeitos às técnicas de *fingerprinting*. Dos mais de 470.000 usuários que participaram de seu projeto público (Panopticlick[1]), 84%

[1] http://panopticlick.eff.org

tiveram seus navegadores identificados.

Formalmente, é importante salientar a diferença entre os termos *Fingerprint* e *Fingerprinting*. De acordo com a RFC 6973 [7], o primeiro é definido como "um conjunto de elementos de informação que define um dispositivo ou uma instância de uma aplicação" e o segundo como "o processo pelo qual um observador ou atacante identifica, de maneira única e com alta probabilidade, um dispositivo ou uma instância de um aplicativo com base em um conjunto de múltiplas informações".

Nesta pesquisa assum-se a visão de que *fingerprinting* é parte de um conjunto amplo de tecnologias e técnicas, também conhecidas como *Device Intelligence*, *Machine Fingerprinting*, *Browser Fingerprinting*, *Web Fingerprinting* ou *Website Fingerprinting*, utilizadas para identificar (ou reidentificar) um usuário ou um dispositivo através de um conjunto de configurações, atributos (tamanho da tela do dispositivo, versões de software instalado, entre muitos outros) e outras características observáveis durante comunicações. Nesta dissertação, os termos *Website Fingerprinting* e *Fingerprinting* serão usados para representar essas técnicas de identificação com foco na Web.

Para melhor ilustrar um *Website Fingerprinting*, a Figura 2.1 exemplifica um usuário visitando um site e tendo seus dados do histórico do navegador coletados por uma empresa de anúncios que consegue utilizar estas informações para verificar detalhes do acesso desse usuário e direcioná-lo para publicidades mais adequadas.

2.1.1 Classificação

De acordo com o W3C (*World Wide Web Consortium*)[2] [8], as técnicas de *Website Fingerprinting* podem ser classificadas em três tipos:

1. **Passiva**: É aquela baseada nas características observáveis no conteúdo de solicitações Web, sem a utilização de qualquer código em execução no lado do cliente. Esse tipo de *fingerprinting* inclui o conjunto de cabeçalhos de solicitação HTTP, endereço IP e outras informações do nível de rede.

[2] http://www.w3c.org

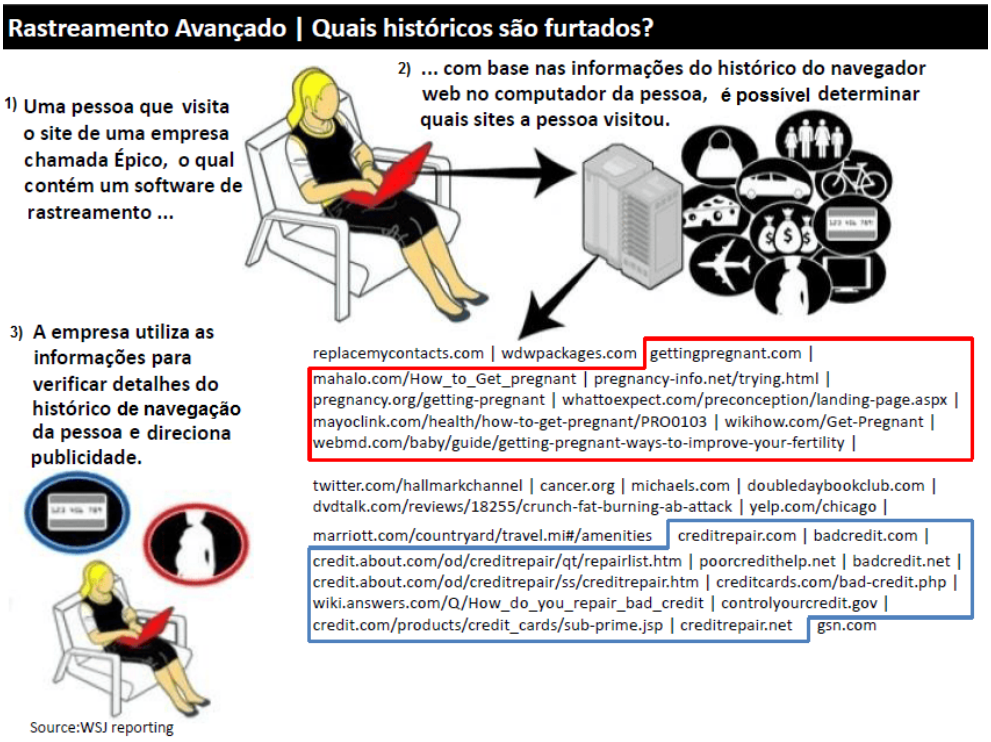

Figura 2.1: Exemplo de *fingerprinting*. Adaptado de The Wall Street Journal [1].

Eventualmente, Cookies também são utilizados. Também é chamada de *Browser Fingerprinting*.

2. **Ativa**: Levam em consideração técnicas onde scripts, via JavaScript (ou outro código), são executados no lado do cliente para observar características adicionais sobre o navegador. Técnicas de *fingerprinting* ativo podem incluir o acesso ao tamanho da janela, enumerar fontes ou plugins, avaliação das características de desempenho ou os padrões de renderização de gráficos, entre outros.

3. **Cookie-like**: Nessa categoria, usuários, *user-agents*[3] e dispositivos, também podem ser (re)identificados por um site que primeiro configura e depois recupera o estado armazenado de um navegador ou dispositivo. A (re)identificação de um usuário ou inferências sobre ele, da mesma forma que os Cookies, permite o gerenciamento de estado para o protocolo HTTP

[3] *user-agent* é um componente do cabeçalho do protocolo HTTP.

(RFC6265 [9]). Essa categoria pode contornar as tentativas do usuário em limitar ou apagar os Cookies armazenados pelo *user-agent*, como demonstrado no trabalho de Kamkar [10].

2.1.2 Uso de Tecnologias Web para *fingerprinting*

Os navegadores tornaram-se as plataformas mais sofisticadas para execução de aplicações Web, assumindo mais funcionalidades do que as tradicionalmente fornecidas pelo sistema operacional. Segundo Mowery e Shacham[11], grande parte deste aumento de sofisticação foi impulsionado por tecnologias e conjuntos de especificações que fornecem a capacidade de desenhar dinamicamente em área da tela (<Canvas>), gráficos tridimensionais (WebGL), armazenamento de dados estruturados do lado do cliente, serviços de geolocalização, capacidade de manipular o histórico e o cache do navegador, reprodução de áudio e vídeo, e muito mais.

Dentre as principais tecnologias utilizadas nos navegadores, destacam-se JavaScript, ActionScript, Canvas, WebGL, CSS e Silverlight, todas importantes para melhorar a dinâmica do conteúdo das páginas Web, porém alvo frequente de *Website Fingerprinting*.

Maiores informações acerca das tecnologias Web relacionadas ao *Website Fingerprinting* podem ser encontradas no minicurso "Device Fingerprinting: Conceitos e Técnicas, Exemplos e Contra-Medidas", publicado no SBSeg 2014 [12], além dos trabalhos de Acar et al. [13], Mowery e Shacham [11] e Englehardt e Narayanan [14] que estão em destaque no Capítulo 3.

É importante ressaltar que esta pesquisa tem como foco Canvas *fingerprinting* que será explorado nas próximas seções.

2.2 HTML5 Canvas

O Canvas é um elemento da HTML5 que fornece uma área da tela que pode ser utilizada via programação. Por meio de JavaScript, Canvas proporciona o acesso a um conjunto completo de funções de desenho, permitindo que gráficos sejam gerados dinamicamente.

A Listagem 2.1 ilustra um código em HTML5 Canvas, cujo resultado é um quadrado vermelho e um texto escrito “Hello World”.

Listagem 2.1: Código exemplo de HTML5 Canvas

```
<script type="text/javascript">
    var Canvas = document.getElementById("Canvas");
    var context = Canvas.getContext("2d");
    context.fillStyle = "rgb(255, 0, 0)";
    context.fillRect(30, 30, 50, 50);
    context.font = "20px serif";
    context.fillStyle = "rgb(0, 0, 255)";
    context.fillText("Hello World", 100, 100);
</script>
```

A Figura 2.2 mostra o fluxo de operações de um Canvas *fingerprinting*.

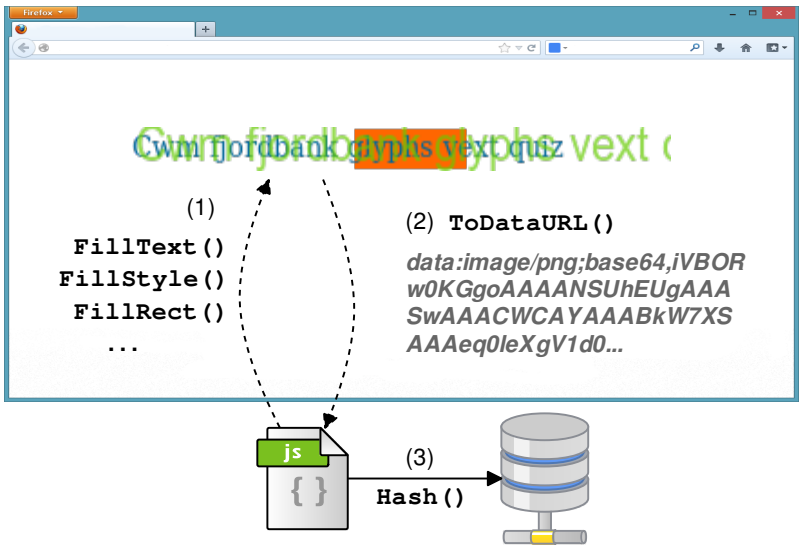

Figura 2.2: Exemplo de *fingerprinting* utilizando Canvas. Fonte: Acar [13].

Quando um usuário visita a página, o script *fingerprinting* primeiro desenha um texto com a fonte e o tamanho de sua escolha e acrescenta cores de fundo (1). Em seguida, o script chama o método *toDataURL*, da API Canvas, para obter os dados de pixel da tela em formato *DataURL* (2), que é basicamente uma representação codificada em Base 64 dos dados de pixel binários. Por fim, o script leva o *hash* dos dados de pixel codificada de texto (3), que serve como *fingerprint* e pode ser combinada com outras propriedades de alta entropia [4] do

[4] A entropia é a medida do grau de desordem, desorganização de um sistema de comunicação,

navegador, como a lista de plugins, a lista de fontes ou a string *user-agent*.

Khademi (2014) [16] destaca que quanto maior a entropia, maior a probabilidade do dispositivo ser identificado de maneira única.

Canvas *Fingerprinting*

Existem trabalhos que mostram a capacidade de Canvas em fornecer uma identificação única rapidamente.

Mowery e Shacham [11] observaram que é possível relacionar o navegador, com maior intimidade, as funcionalidades do hardware e do sistema operacional. Os autores desenvolveram uma técnica que quando um usuário visita um site que utiliza *fingerprinting*, o navegador desenha uma linha oculta de texto ou de gráfico 3D, a qual é convertida em um sinal digital. Desta maneira, diferentes placas gráficas instaladas no computador do cliente, juntamente com os diferentes drivers, causam variações em *tokens* digitais. O *token* pode ser armazenado e compartilhado com empresas de publicidade para identificar os usuários quando eles visitam sites afiliados. Um perfil pode ser criado como atividade de navegação de um usuário, permitindo que anunciantes direcionem sua publicidade de acordo com as preferências do usuário.

Já o estudo de Kirk [17] relata a utilização de código para *fingerprinting*, com base na tecnologia Canvas, que estava em uso no início de 2014 em mais ou menos 5000 sites populares, sem qualquer tipo de conhecimento para seus usuários. Entretanto, nem todos os locais observados com *fingerprinting* faziam o compartilhamento de conteúdo para empresas de publicidade. Ele ressalta ainda que as empresas europeias estão à procura de novas maneiras de entregar publicidade segmentada aos usuários, afastando-se dos Cookies.

Um exemplo prático desse tipo de *fingerprinting* pode ser vislumbrado no site *propublica.org*, que gera um *hash* do dispositivo (Figura 2.3).

A ideia exemplificada na Figura 2.3 é que mesmo a menor alteração em um pixel (um ponto na imagem) pode criar um Código Identificador (ID) totalmente novo. Diferentes computadores e navegadores Web podem desenhar a imagem de forma diferente, resultando em um ID que é semi-exclusivo para um usuário.

a falta de previsibilidade, o que resulta na incerteza, ou seja, na falta de conhecimento de um determinado assunto (Duarte, 2010 [15]).

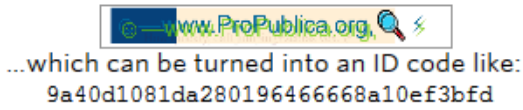

Figura 2.3: Exemplo de Site para Testar o Dispositivo do Usuário em relação ao Canvas *Fingerprinting*. Fonte: Propublica.org

Os IDs podem ser usados para acompanhar os usuários de um site para outro - mesmo quando os cookies estão desativados no navegador Web de um usuário.

2.3 Recuperação de Informação

Recuperação de Informação (RI) é a área da computação que objetiva lidar com documentos para armazená-los e recuperar automaticamente a informação relacionada a estes. Baeza-Yates e Ribeiro-Neto [18] deixam claro que a RI trata da representação, armazenamento, organização e acesso a itens de informação, como: documentos, páginas Web, catálogos on-line, registros estruturados/semi-estruturados, objetos multimídia, entre outros.

Esta representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse. Deste modo, o seu principal objetivo é recuperar informações, contidas em documentos, que possam ser relevantes e úteis para os usuários. As demais subseções destacam a taxonomia dos Modelos de RI (tipos de modelos utilizados para a recuperação da informação) e o Modelo Vetorial adotado nesta dissertação.

2.3.1 Modelos de RI

Os modelos de RI fornecem princípios para a construção da função de ordenação e praticamente são baseados por textos usados para o ranqueamento. No entanto, tratando-se de informações Web, também deve-se verificar links e objetos multimídia que não são codificados da mesma forma que os textos. As imagens são codificadas como mapa de pixels, vídeos como fluxos (*streams*[5]). Por isso, os modelos de RI podem ser classificados como os baseados em: texto, em links

[5]Stream é o fluxo de dados ou de conteúdo multimídia

e objetos multimídia. Os modelos clássicos do tipo texto não estruturado mais conhecidos são explicitados por Baeza-Yates e Ribeiro-Neto [18]:

- **Modelo Booleano**: baseado na composição de um conjunto de documentos e operações clássicas da teoria de conjuntos;
- **Modelo Vetorial**: no qual documentos e consultas são representados como vetores em um espaço n-dimensional utilizando operações da álgebra linear aplicáveis aos vetores;
- **Modelo Probabilístico**: baseado na teoria das probabilidades realizando assim, representações para documentos e consultas.

Nesta pesquisa, o modelo para verificação dos scripts Canvas *fingerprinting* fora o Modelo vetorial com a finalidade de identifica-los, visto que neste modelo os documentos são recuperados de modo decrescente fornecendo uma resposta mais precisa do que nos demais modelos anteriormente mencionados. Para fins de entendimento, a Figura 2.4 apresenta um script, o modelo vetorial, bem como documentos, consultas, termos e pesos.

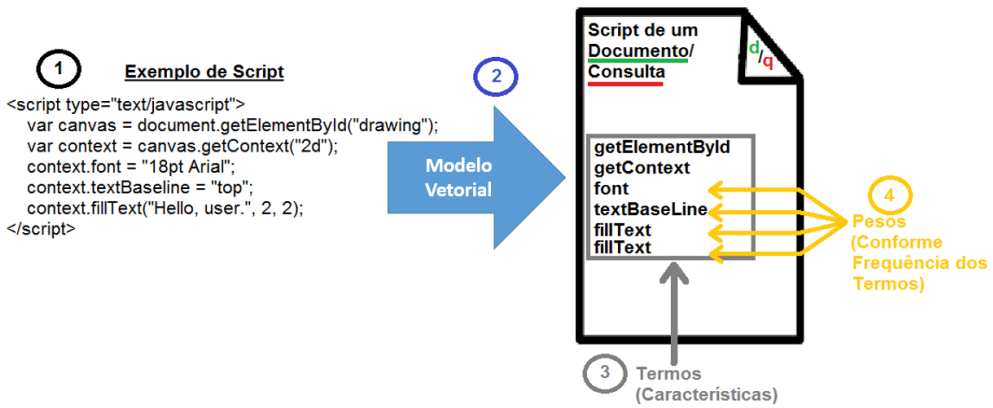

Figura 2.4: Exemplo de como o Método Vetorial trata os Documentos, Consultas, Termos e Pesos no espaço n-dimensional.

A Figura 2.4 apresenta um trecho de um script qualquer que sofrerá à ação do modelo vetorial. Como resultado, o script (documento/consulta no modelo vetorial) é transformado em termos, que são as características (propriedades e

métodos Canvas *fingerprinting*. Em seguida, os pesos de cada termo são calculados de acordo com sua frequência de aparição. Com base nesses pesos, ocorre o cálculo da similaridade e o ranqueamento dos scripts mais semelhantes entre si.

O Modelo vetorial está melhor detalhado na seção 2.3.2.

2.3.2 Modelo Vetorial

O modelo vetorial, também conhecido como modelo espaço vetorial, é um modelo onde documentos e consultas são vistos como características num espaço vetorial n-dimensional, sendo a distância vetorial usada como medida de similaridade[6]. De acordo com Oliveira [19], "o modelo vetorial visa recuperar informação de forma simples e eficiente. Nesse modelo é possível obter documentos que respondam parcialmente a uma expressão de busca".

O modelo vetorial forma vetores no espaço euclidiano, onde cada termo corresponde a um eixo no espaço n-dimensional, com base no peso do termo em relação ao documento inteiro (Ramiro et al., 2005 apud Chaves, 2014 [20]). Assim, são listados os documentos e os índices obtidos em todos os documentos analisados e é aplicada a ponderação de cada índice em relação a cada documento.

Para esta pesquisa, o modelo vetorial é de suma importância, pois tem a finalidade de realizar uma verificação na similaridade entre os documentos (bases de dados com scripts) e as consultas contendo Canvas. As próximas subseções destacam de maneira formal os termos: documento, consulta, termo, peso e similaridade, que são fundamentais para o entendimento desta pesquisa.

Documento e Consulta

Um documento d_j e uma consulta de usuário q são representados como vetores com t dimensões. O modelo vetorial calcula o grau de similaridade do documento d_j em relação à consulta q sob forma de correlação entre os vetores d_j e q, podendo ser quantificada, por exemplo, pelo cosseno do ângulo entre esses dois vetores (Baeza-Yates e Ribeiro-Neto) [18].

[6]Similaridade é a distância entre características no espaço vetorial e pode ser calculada pela comparação de seus vetores, usando uma medida tal como o cosseno.

A Figura 2.5 destaca um exemplo de como o modelo vetorial faz uso de um vetor de documentos, posicionando um documento no espaço vetorial de três dimensões, cada uma representando um índice. Cada documento é visto como termos no espaço n-dimensional através dos eixos: x=4, y=5 e z=3. O ponto de interseção entre estes termos é representado pelo documento DOC = (5, 4, 3).

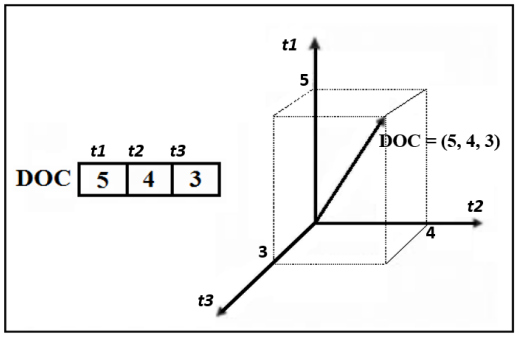

Figura 2.5: Exemplo de como o Método Vetorial trata os Documentos no espaço n-dimensional. Fonte: Adaptado de Ramiro et al, 2005 apud Chaves, 2014 [20]

Em outra perspectiva, a Figura 2.6 ilustra como os termos são tratados no espaço n-dimensional. Em linhas gerais, cada termo tem um peso, que é calculado através de uma matriz de pesos gerada com base na frequência de aparição dos termos. Desta forma, o cálculo da similaridade (que será melhor explicado nas próximas seções) ocorre comparando os termos entre os documentos no espaço n-dimensional.

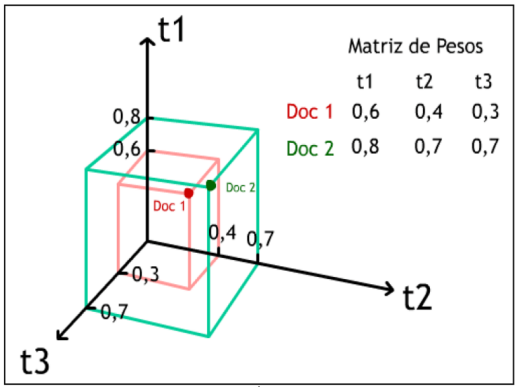

Figura 2.6: Exemplo de como o Método Vetorial trata os Termos e Pesos no espaço n-dimensional. Fonte: Chaves, 2014 [20]

Um documento nesta pesquisa pode ser exemplificado como um dos scripts de uma das bases de dados aqui utilizadas. Já a consulta pode ser exemplificada como um dos scripts separados para servir de comparação com o documento da base de dados. Um fator importante para o método vetorial, no que tange a comparação anteriormente mencionada, é o termo a ser analisado e do peso que este termo tem em relação à consulta, os quais serão melhor elucidados na subseção a seguir.

Na próxima seção é elucidada a explicação de como o método vetorial trata os termos e os pesos no espaço n-dimensional.

Termos e Peso

Para o modelo vetorial, o peso $w_{i,j}$ associado ao par termo-documento (k_i,d_j) é não negativo e não binário. Os termos de indexação são todos considerados mutuamente independentes e são representados por vetores unitários em um espaço com t dimensões. Tanto documento d_j, quanto consultas q são vetores com t dimensões dados por:

$$w_{i,j} = tf_{i,j} \times idf_i$$
$$w_{i,q} = tf_{i,q} \times idf_i$$

onde, w_i,q é o peso associado ao par termo-consulta (k_i,q), com w_i,$q >= 0$.

O idf_i (*inverse document frequency*) representa a importância do termo t_i para a coleção de documentos e pode ser calculado por meio da Equação 2.1:

$$idf_i = \log\left(\frac{N}{n_i}\right) \tag{2.1}$$

onde, N é o número de documentos da coleção e ni é o número de documentos onde ocorre o termo t_i (Baeza-Yates e Ribeiro-Neto) [18].

Cada documento é representado como um vetor de termos e cada termo possui um valor associado que indica o grau de importância (peso) deste em um

determinado documento.

Neste trabalho, os termos são as características (propriedades e métodos) Canvas *fingerprinting* utilizados para averiguar nas bases de dados (documentos) e nas consultas as características existentes nestes. Já os pesos podem ser entendidos como a quantidade de vezes que estas características aparecem nos documentos e nas consultas.

Similaridade

O modelo vetorial calcula o grau de similaridade do documento d_j, em relação à consulta q sob forma da correlação entre d_j e q. Essa correlação pode ser calculada por meio de qualquer medida da relação entre vetores. Neste caso, a medida de relação adotada é a similaridade entre cossenos (*cosine similarity*), isto é, o valor de cosseno do ângulo formado entre eles. Desta forma, a similaridade entre o documento d_j e a consulta q é calculada através da similaridade do cosseno através da equação a seguir:

$$sim(d_j,q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2.2}$$

Sendo $w_{i,j} >= 0$ e $w_{i,q} >= 0$, a similaridade $sim(d_j,q)$ varia entre 0 a 1 (Baeza-Yates e Ribeiro-Neto) [18]. Desta maneira, o modelo vetorial é capaz de ordenar os documentos de acordo com o grau de similaridade de cada documento com a consulta realizada pelo usuário. Um documento pode ser recuperado mesmo se ele satisfaz a consulta apenas parcialmente (a exemplo disso, das 41 propriedades/métodos de Canvas *fingerprinting* utilizados, assumindo que somente duas sejam encontradas na comparação entre um documento e uma consulta, será apresentado um grau de similaridade entre este documento e esta consulta mesmo assim).

Este trabalho utiliza o modelo vetorial tradicional, o qual serve para transformar os scripts da coleção em um conjunto de termos. Deste modo, podem ser utilizadas técnicas de RI para calcular a similaridade entre os scripts.

2.4 Detecção de *Knee Points*

Esta pesquisa emprega o conceito de detecção de *Knee Points* (em tradução livre, pontos de joelho), definida por Satopää et al. [21], com a finalidade de encontrar um ponto em uma função na qual não haverá tanta variação de resultados e realizar um corte nos resultados de similaridade naqueles em que o algoritmo não considerar relevantes. A ideia de detecção de *Knee Points* baseia-se na definição matemática de curvatura para uma função contínua, na qual, para qualquer função contínua f existe um padrão de forma fechada $K_f(x)$ que define a curvatura de f em qualquer ponto como uma função de sua primeira e segunda derivada:

$$K_f(x) = \frac{f"(x)}{(1 + f'(x)^2)^{1.5}} \tag{2.3}$$

A Figura 2.7 ilustra o ponto de curvatura máxima que é adaptado aos operadores para selecionar um joelho, uma vez que a curvatura é uma medida matemática de quanto uma função difere de uma linha reta. Como resultado, a curvatura máxima capta o fim do efeito de nivelamento que os operadores usam para identificar os joelhos. Importante, que diferentemente de outras definições comuns, a curvatura é aplicação e **(i)** não depende da relação entre parâmetros e desempenho do sistema, ou **(ii)** exige uma definição de limites específicos do sistema.

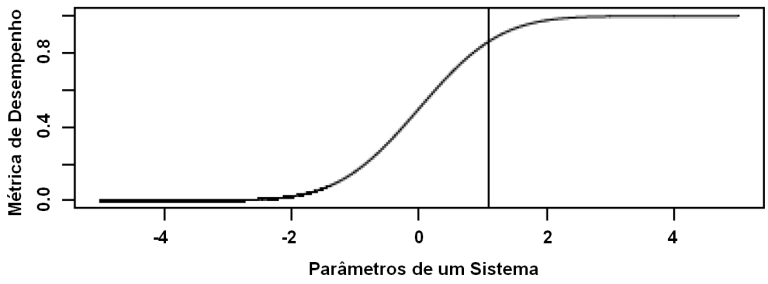

Figura 2.7: Gráfico de uma distribuição Gaussiana padrão com média = 0 e desvio padrão = 1. Barra vertical indica ponto de curvatura máxima. O ponto de inflexão desta curva ocorre em x = 0. Fonte: Satopää et al. [21].

Na Figura 2.7, o eixo y representa alguma métrica de desempenho, o eixo x representa um parâmetro de um sistema e a barra vertical representa o ponto de curvatura máxima. O máximo da primeira derivada é o ponto de inflexão da curva, que ocorre em x = 0 na Figura 2.7. O ponto de inflexão não é representativo do Joelho já que o desempenho continua a melhorar significativamente além disso. Em vez disso, o ponto de inflexão apenas captura a taxa de aumento de desempenho ao atingir um valor máximo. Em contraste, a definição de curvatura corresponde precisamente ao conceito de joelho Satopää et al. [21].

O exemplo a seguir elucida o uso de *Knee Points* neste trabalho. Ao realizar a comparação entre uma das 100 consultas de Canvas *fingerprinting* com os documentos (da base de Canvas *fingerprinting* que possui 8.000 scripts), supõem-se que são obtidos mais de 4.000 scripts similares a esta consulta. Neste caso, o algoritmo *Knee Points* irá verificar o ponto da curvatura da função em que não há tanta representatividade (devido a invariância dos resultados de similaridade) e irá eliminar estas comparações, resultando, por exemplo, em 2.000 scripts similares. No capítulo 4 seção 4.1.5 constam maiores informações sobre as 100 consultas mencionadas anteriormente.

Satopää et al. [21] destacam que detecção de *Knee Points* é um processo inerentemente heurístico que independe da aplicação, ou seja, em uma definição coerente ele pode ser aplicável a qualquer sistema. Contudo, existe uma grande dificuldade em defini-lo formalmente, pois este pode ser "bom o suficiente" em um sistema, mas pode não ser "bom o suficiente" em outro. Para o método proposto nesta pesquisa, o *Knee Points* obteve um desempenho bom o suficiente para diminuir consideravelmente a quantidade de amostras.

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta uma relação de trabalhos referentes a *Website Fingerprinting* encontrados na literatura. Como forma de melhor descrevê-los, os trabalhos são divididos em duas seções. A primeira destaca os principais trabalhos que comprovam a existência de técnica de *Website Fingerprinting* para identificar um usuário/dispositivo. Já a segunda apresenta os trabalhos que abordam Canvas *Fingerprinting*, foco desta pesquisa. Além disso, ao final deste capítulo é apresentada uma discussão sobre esses trabalhos.

3.1 Existe *Website Fingerprinting*?

O primeiro trabalho a de fato provar a existência de *Website Fingerprinting* foi feito por **Peter Eckersley** em 2010 [6]. O artigo aborda sua experiência ao verificar o quanto a configuração de um navegador é única e utilizável como possível mecanismo de rastreamento de usuários na Internet. O foco da pesquisa foi investigar o grau no qual os navegadores modernos estão sujeitos a *Website Fingerprinting*, através da análise de versões e informações de configurações que podem ser coletadas com ou sem o consentimento prévio dos usuários. Eckersley foi capaz de combinar todas as informações obtidas e gerar um identificador (ID) único para cada dispositivo. A pesquisa fez uso de uma página Web (**Panopticlick** [1]) na qual um algoritmo de *fingerprinting* foi aplicado. Assim, cada navegador que

[1] https://panopticlick.eff.org

a acessava tinha um identificador gerado.

Dentre os 470.161 acessos ao site (os usuários foram informados do propósito da pesquisa ao visitarem à página), Eckersley obteve 83,6% de identificações únicas, com apenas 5,3% de navegadores/usuários conseguindo se manter no anonimato. Também foi observado que 94,2% dos navegadores que possuíam adobe Flash ou tinham habilitado o suporte Java foram identificados unicamente. Apenas 1,0% dos navegadores com Flash ou Java mantiveram o anonimato. O autor destaca que as identificações únicas não são definitivas, visto que constantemente ocorrem mudanças nos navegadores como: atualização de versões ou de plug-ins, desabilitação de cookies, instalação de fontes. Porém, o fato alarmante é que seu algoritmo foi capaz de identificar todas as mudanças ocorridas, chegando a um percentual de 99,1% de acerto, com uma taxa de falso positivo de apenas 0,87%.

Jang et al. [22] apresentaram um estudo empírico da prevalência de fluxos de informações que tendem a violar a privacidade do usuário através de implementações de código JavaScript. Para tanto, os autores implementaram uma nova versão do motor Javascript dentro do navegador Chrome e analisaram mais de 50.000 sites do Alexa.com.

Como resultado, eles conseguiram identificar instâncias de geolocalização bem como diversas agências de publicidade que utilizam informações contidas nos Cookies. Isso possibilitou: (i) detectar sites que utilizam as informações do histórico como mecanismo de *fingerprinting*; e (ii) verificar que sites populares (Microsoft e fingtonpost.com, por exemplo) possuem em sua infraestrutura mecanismos para rastrear os movimentos e cliques dos mouses dos usuários.

Boda et al. [23] propuseram e avaliaram um método de *fingerprinting* independente do navegador. Para tanto, criaram um identificador único de dados de um navegador específico usando algoritmos JavaScript do lado do servidor. O trabalho discute melhorias no algoritmo utilizado no Panopticlick de **Peter Eckersley** [6] na identificação de *fingerprinting*, com base em um conjunto de mil registros coletados por meio de sites publicamente acessíveis. Os pesquisadores demonstram como uma parte do endereço IP, um conjunto de fontes, o fuso horário e a resolução de tela são suficientes para identificar a maioria dos usuários dos cinco maiores navegadores utilizados atualmente.

Como resultados, os autores fazem: (i) uma correlação com as listas de fontes

e UserAgent, proporcionando a identificação única dos SO Windows e Mac; (ii) desenvolveram um algoritmo semelhante ao do Panopticlick, porém capaz de controlar eficientemente mudanças no endereço IP (por exemplo, quando um usuário se reconecta/muda de rede) e distinguir entre diferentes PCs atrás de um NAT (*Network Address Translation*); e também mais resistente à atualizações no computador, no navegador, comutação de browsers, (des)instalação de plug-ins, e exclusão dos UserAgent.

Yen et. al. [24] realizaram um estudo em grande escala para quantificar a informação que pode ser revelada por um servidor, fato que pode implicar na privacidade e segurança das informações do usuário. Os autores analisaram um conjunto de dados anônimos (endereços IP, Cookies, IDs de login do usuário, entre outros) coletados utilizando o serviço de Webmail do Hotmail e do motor de busca Bing. As análise mostraram que mesmo um usuário que realiza constantemente a limpeza de seus Cookies ou que utiliza a navegação privada pode ser rastreado e, que ataques maliciosos podem ser camuflados por mais de 75.000 contas *bot* que encaminham Cookies para locais distribuídos, através da análise de One-time Cookies[2].

Como resultado final, o estudo mostrou que entre 60% a 70% dos valores do *user-agent* podem identificar com precisão os visitantes. Também mostrou que essa característica, quando combinada com o endereço IP, tem uma entropia[3] melhor na classificação de usuários do que a combinação entre plug-ins, resolução de tela, fontes e fuso horário. Além do mais, o uso de One-time Cookies permitiu o identificar 88,00% dos usuários que voltam a um serviço (entre estes usuários, 33% fizeram um esforço para preservar a sua privacidade).

Acar et al. [25] desenvolveram um framework (FPdetective) para detectar códigos de *fingerprinting*. A ferramenta é composta por um *Crawler* (para direcionar a coleta das páginas), um *Parser* (para extrair dados dos scripts e armazená-los em um banco de dados), um proxy (para analisar e extrair arquivos Flash) e um decompilador (que busca por chamadas de funções de *fingerprinting*.

FPDetective foi capaz de detectar 404 sites, entre os Top 1 milhão do Alexa.com,

[2]One-time Cookies - Cookies que parecem ser anônimos

[3]Entropia caracteriza a impureza de uma coleção arbitrária de exemplos, ou seja, é a medição da Homogeneidade dos Exemplos (Koerich [**?**]).

que se utilizam de JavaScript para realizar *fingerprinting*. Além disso, foram encontrados 145 sites que empregavam Flash para *fingerprinting*, sendo que com maior capacidade de identificação (enumeração de fonte, detecção de proxy e compatibilidade com navegadores).

Olejnik et al. [26] realizaram uma pesquisa em grande escala, na qual cerca de 368.284 usuários tiveram seus históricos de navegação à Internet detectados. Para tanto, foram utilizados experimentos com base em CSS.

Os resultados mostram que para a maioria dos usuários (69%), o histórico de navegação era único. Com isso, foi possível detectar pelo menos quatro sites exclusivamente por seus históricos em cerca de 97% dos casos. Observou-se uma taxa significativa de estabilidade na *fingerprinting* do histórico do navegador, uma vez que: (i) ao repetir os testes, 38% das *fingerprinting* são idênticas ao longo do tempo, e as diferentes foram correlacionadas com conteúdo dos históricos originais, indicando preferências de navegação estáticas; (ii) o teste com apenas 50 páginas foi suficiente para ter uma acurácia de 42% dos usuários no banco de dados, aumentando para 70%, com 500 páginas Web.

3.2 Identificação de usuários/dispositivos via Canvas *Fingerprinting*

Englehardt e Narayanan [14] realizaram uma análise dos sites que fazem parte dos tops 1 milhão de sites da base Alexa.com. Os autores utilizaram uma plataforma denominada OpenWPM, a qual vasculhou os sites listados com o objetivo de analisar quais deles rastreiam seus usuários, seja por cookies (rastreamento por estado) ou por *fingerprinting* (métodos sem estados).

Diferente de trabalhos anteriores, o uso da plataforma OpenWPM permitiu que a busca (*crawling*) fosse realizada automatizando navegadores robustos (isto é, aqueles navegadores que são usados comumente como Firefox e Chrome), através da ferramenta Sellenium, o que pode ser visto como uma vantagem. A intenção dos autores foi mostrar a importância de recursos como Canvas, conteúdos dinâmicos, e outros, visto que muitos sites funcionam corretamente apenas em navegadores mais robustos (como o Chrome, e o Firefox), diferentemente dos

navegadores como o PhantomJS, usado por outros estudos.

Além disso, a pesquisa destaca novos métodos de rastreamento e de persistência da identificação, como o da API *AudioContext*, API *Battery* e *WebRTC* (capaz de descobrir o IP privado por trás de proxy e/ou NAT). Já em relação aos métodos de persistência, o estudo mostra que muitos sites utilizam o cookie sync (compartilhamento de cookies entre terceiros), permitindo que, mesmo que um usuário apague todos os cookies, os mesmos sejam restaurados a posteriori. Porém, o mais alarmante exposto na pesquisa dos autores está no fato de que alguns métodos do HTML5 permitem restaurar esses identificadores de forma mais transparente do ponto de vista do usuário, com o uso do *localStorage* do HTML5.

Como resultado em relação a Canvas, os autores destacaram que foram detectados 14.371 sites, da base de 1 milhão (cerca de 1,6 %), contendo as propriedades e métodos como *toDataURL*, *getImageData*, *save* e *restore*.

Laperdrix et al. [27] realizaram um estudo sobre os atributos mais relevantes para extrair o *fingerprinting* do dispositivo do usuário e analisar a eficácia da técnica tanto em dispositivos móveis, quanto em computadores. Os autores fizeram um estudo sobre as tecnologias Web, demonstrando a perspectiva negativa e positiva de cada uma delas, com a finalidade de expor como as tecnologias podem ser utilizadas para invadir a privacidade do usuário.

Deste modo, criaram um script de *fingerprinting* que faz uso de 10 atributos utilizados no trabalho de Eckersley [6], com 7 atributos adicionais que utilizam novos atributos provenientes dos mais recentes avanços das tecnologia Web (Java Script, HTML5, Canvas e etc). Os autores criaram uma base de *fingerprinting* a partir dos experimentos realizados com o script hospedado no site AmIUnique.org. Com posse da base contendo 118.934 dados coletados, os autores identificaram de maneira única 89.4% dos dispositivos analisados.

É importante destacar que não foram utilizados somente atributos da tecnologia Canvas, mas sim um total de 17 atributos, porém, os autores explicitam que o HTML5 Canvas está no top 5 dos atributos mais influentes. Com isso, ao final da pesquisa, os autores descreveram os impactos dos atributos coletados, além de fazer uma comparação com o trabalho do Eckersley [6] e de prever como o desaparecimento de algumas APIs (Flash por exemplo) podem impactar na

unicidade do *fingeprinting* de um dispositivo.

Ximenes et al. [4] desenvolveram um mecanismo capaz de identificar unicamente um navegador Web através do uso do elemento Canvas. Este mecanismo mostra-se diferente dos já existentes por ser resistente às técnicas de contramedidas que bloqueiam o Canvas e/ou alteram seu comportamento.

Em seus experimentos, os autores pretendiam verificar se poderiam gerar objetos <Canvas> com imagens de alta e baixa entropia simultaneamente em um mesmo navegador. Para tanto, desenvolveram um protótipo (TARP *Fingerprinting*) e o aplicaram a um conjunto de usuários da Web. No protótipo foram utilizados três grupos distintos de instruções gráficas HTML 5 tipo Canvas. O primeiro denominado de grupo L (modelado para assinaturas de baixa entropia) foi construído a partir da sobreposição de duas formas gráficas simples. O segundo, denominado de grupo H (modelado para produzir assinaturas de alta entropia) foi construído a partir da biblioteca *fingeterprint2.js*, por possuir uma diversidade de variações de elementos gráficos Canvas. E por último, o de instruções grupo M (modificação das instruções do grupo H) foi construído para incrementar a entropia das assinaturas geradas.

Com isso, o protótipo gera imagens de baixa entropia, por meio da execução das instruções do grupo L, e imagens de alta entropia ,através da execução das instruções dos grupos H e M. Para diminuir o consumo de armazenamento em memória, as assinaturas foram calculadas via *hash* MD5 do arquivo binário de cada uma das imagens geradas. Desta maneira, o protótipo gera uma assinatura única de baixa entropia (que pode ser usada como marca d'água) e duas assinaturas de alta entropia (usadas para compor um identificador distinto por instância).

Os dados da pesquisa foram coletados via site de uma universidade brasileira, onde todas as visitas geraram uma instância do TARP *fingerprinting*. Foram coletadas 64.086 assinaturas (cada assinatura equivale a uma instância do TARP), na qual por meio do cálculo da entropia era verificado o uso ou não de contramedidas. Para tanto, os autores utilizaram a entropia ponderada e a de Shannon, resultando em um conjunto de três valores, via TARP *fingerprinting*, respectivamente: TARP L = 2,43; TARP M = 7,68 e TARP H = 7,86. E a posteriori realizaram uma composição (união) dos Tarps, eles obtiveram: TARP LM =

7,86; TARP HM = 7,91 e TARP LHM = 7,91. Destacando alta entropia em relação as comparações com os experimentos de Mowery & Shacham [11] e de Laperdrix [27].

Bursztein et al. [28] desenvolveram e avaliaram uma técnica de *Website Fingerprinting* com a finalidade de detectar e prevenir ataques automatizados de clientes maliciosos em lojas de aplicativos. O estudo objetivava distinguir os clientes autênticos dos emulados/automatizados. Os autores, diferentemente da maioria dos demais trabalhos os quais buscaram identificar unicamente um dispositivo, fizeram a identificação de classes de dispositivos, através da unificação de atributos do navegador Web e do sistema operacional. A ideia central ocorre por desafios elaborados a partir da tecnologia HTML 5 Canvas.

Esses desafios baseiam-se em quatro operações de desenhos, que são denominadas de primitivas gráficas, conforme destaque: *arc()*, *strokeText()*, *bezierCurveTo()*, *quadraticCurveTo()*. Estas primitivas oportunizam o cálculo de ângulos e verificam o formato de fontes. Após este teste, cada elemento gráfico criado passa por um processo de customização, por meio das operações: *createRadialGradient()*, *shadowBlur()*, *shadowColor()*, com vistas a obtenção do máximo de entropia possível dos elementos gráficos.

Em seus experimentos os autores parametrizaram as operações para a realização dos desafios de modo aleatório fazendo dois tipos de experiências, uma controlada e outra aberta. O primeiro dispôs de uma amostra contendo 272,198 dispositivos. Já o segundo em ambiente aberto contou com a participação anônima de 52 milhões de usuários em um ambiente de produção. Em ambos os casos, a ferramenta conseguiu fazer a distinção das classes dos dispositivos com 100% de acurácia, aplicando-se para todas as combinações de navegadores e sistemas operacionais (sejam estes móveis ou desktops).

Nakibly, Shelef e Yudilevich [29] apresentaram técnicas de *Device Fingerprinting* que utilizam o HTML5 e também as características dependentes não somente do software, como a maioria das pesquisas, mas também do hardware. Os autores propuseram um script que faz o uso da GPU e do Canvas para obter o *fingerprint* do dispositivo do usuário.

Eles desenvolveram o site fingerprintme.herobo.com para realizar seus experimentos e coleta de dados. O site exibe gráficos em três fases (1 - verifica a

frequência de relógio; 2 - verifica o número de núcleos e 3 - verifica outros parâmetros que afetam o desempenho da GPU), onde cada fase durou um período de 15 segundos e as medições do número de quadros renderizados foram realizadas em três intervalos de 5 segundos. A partir disso, foi traçado um histograma para cada fase das medições, sendo que cada fase corresponde a um compartimento com largura de 5 quadros. Este processo de renderização deu-se por meio da API Canvas *requestAnimationFrame*, com base em computadores dos tipos desktop e laptop, dos quais obtiveram-se a identificação de 130 dispositivos diferentes, onde 34 destes foram "fingerprintados" mais de uma vez.

Como resultado da pesquisa, para cada uma das fases mencionadas anteriormente os autores efetuaram o cálculo da entropia, obtendo-se os seguintes valores: Fase 1 = 2.8; Fase 2 = 5.03 e na Fase 3 = 5.14.

Acar et al. [13] Instrumentaram um navegador Web com a ferramenta de automação Selenium, com vistas a investigar os 100.000 sites mais populares do Alexa.com, os quais pudessem possuir scripts para rastrear usuários. Neste estudo, eles constataram que o *Website Fingerprinting* é empregando via HTML5 Canvas como mecanismo de rastreamento.

Os autores, utilizando uma versão do Firefox com o código fonte modificado, conseguiram registrar todas as chamadas de funções que poderiam ser utilizadas no Canvas *fingerprinting*. Inicialmente, eles elaboraram formas para identificar o Canvas *fingerprinting* e desenvolveram um *crawler* exploratório. Com isso, obtiveram um método automatizado baseado nos estudos de Mowery e Shacham [11]. Em seguida, o estudo empregou métodos analíticos reportados na literatura sobre Canvas *fingerprinting*. Dentre os métodos utilizados para detectar o Canvas *fingerprinting*, destacam-se *toDataURL*, *fillText* e *strokeText*, onde o primeiro é empregado para ler o Canvas e os demais são empregados com a finalidade de desenhar textos.

Como forma de evitar resultados falsos positivos, os autores consideraram pontos como: (i) deve haver chamadas para ambos os métodos *toDataURL* e *fillText*, porém estas chamadas devem ser originárias de um mesmo domínio (URL); (ii) as imagens de tela lidas pelo *script* devem conter mais de uma cor e seu tamanho deve ser superior a 16x16 pixels; (iii) a imagem não deve ser solicitada em formato de compressão de perdas, como JPEG por exemplo.

O resultado mostrou que dos 100.000 sites analisados, 5.5% utilizam a tecnologia HTML5 Canvas para realizar *Website Fingerprinting*, sendo 95% dos scripts de Canvas pertencentes a um único provedor, neste caso o *addthis.com*, e os demais 5% restantes pertencem a outros 20 provedores (11 de companhias terceirizadas e 9 desconhecidos).

3.2.1 Discussão

Após a explanação sobre os trabalhos, percebe-se que, embora exista o combate ao *Website Fingerpriting*, as pesquisas pautaram-se em provar a existência do problema, desenvolvendo, em sua maioria, algum site para realizar esta identificação dos usuários/dispositivos. Além disso, a maioria das pesquisas mencionadas neste capítulo tem uma abordagem on-line e não faz uso de scripts. Tal fato fica claro pela dificuldade em afirmar que um site e seus scripts, de fato, fazem *fingerpriting* e são maliciosos para a privacidade do usuário.

No que tange Canvas *fingerprinting*, a Tabela 3.1 sumariza os trabalhos apresentados e os discute a seguir.

Tabela 3.1: Discussão

Tabela de Discussão				
Autores	**Método de Detecção**	**Propriedades Canvas**	**Protótipo**	**Fonte de Dados**
Englehardt e Narayanan (2016)[14]	Ferramenta OpenWPM	Canvas (toDataURL, getImageData, save e restore), AudioContext, Battery e Web RTC	Sim	OpenWPM
Laperdrix (2016)[27]	*	Canvas (fillText, fillStyle, fillRect, toDataURL, strokeText, globalCompositeOperation, lineTo, arc, canvas.Text, getImageData), WegGL (Vendor e Renderer), Plataform (UserAgent), Do Not Track e Ad blocker	Não	AmIUnique
Ximenes et al.(2016)[4]	Entropia Ponderada e de Shannon	Canvas (toDataURL)	Sim	Tarp FP
Bursztein et al. (2016)[28]	Máxima Entropia	Canvas (arc, strokeText, bezierCurveTo, quadraticCurveTo, createRadialGradient, shadowBlur, shadowColor)	Sim	Picasso
Nakibly et al. (2015)[29]	fingerprintingme.herobo.com	Canvas (requestAnimationFrame), GPS, GPU, Sensor de Presença, Microfone, Caixa de Som	Sim	fingerprintme
Acar et al. (2014)[13]	*	Canvas (toDataURL, fillText, strokeText)	Não	alexa.com
Avaliando Canvas FP	Método Vetorial e *Knee points*	Canvas (Propriedades e Métodos) Vide Tabelas 4.4 e 4.5 do Capítulo 4, seção 4.3	Não	Diversas

Em relação às características (Propriedades e Métodos Canvas *fingerprinting*), a Tabela 3.1 destaca que a maioria das propriedades Canvas utilizadas nas pesquisas são: *filltext*, *fillRect*, *fillStyle*, *toDataURL*, *strokeText*, *arc*, *quadraticCurveTo*, *globalCompositeOperation*, *lineTo*, *canvas.Text*, *getImageData* e *requestAnimationFrame*. Percebe-se que algumas propriedades do Canvas *fingerprinting* são bastante recorrentes, mas outras são mais raras, fato que proporciona uma atenção maior a estas com um número de incidências menor. Por exemplo, as

propriedades *toDataURL* e *getImageData*, presentes nos trabalhos de Englehardt e Narayanan [14], Laperdrix et al. [27], Ximenes et al. [4] e Acar et al. [13], são elencadas no trabalho de Saraiva [30] como características de alta periculosidade, por evidenciar a prevalência maior da existência ou não do Canvas *fingerprinting*.

Contudo, nesta pesquisa não são utilizadas somente essas 17 propriedades/-métodos mencionados. Outras 24 propriedades/métodos da API HTML5 Canvas foram incorporadas para uma análise mais abrangente sobre o Canvas *fingerprinting*.

Já sobre a classificação e o ranqueamento, é importante salientar que não existem trabalhos que empregam um método para predizer os níveis de similaridade. Tudo leva a crer que esta pesquisa é a primeira a fazer isso. Vale destacar que a abordagem proposta tem a ver com a questão do ranqueamento, o qual irá dispor dos Top 5 scripts, com vistas a mostrar os cinco scripts mais similares as consultas exploradas nesta pesquisa.

Sobre o desenvolvimento de protótipo, a abordagem proposta não possui um protótipo formal, visto que este não é o foco desta pesquisa. Entretanto, ela poderá ser incluída em um plugin ou aplicação, proporcionando a averiguação de um determinado site em relação ao seu nível de similaridade.

Acerca das bases de dados utilizadas, os trabalhos na Tabela 3.1 utilizam-se da base de dados do Alexa.com ou elas mesmas coletaram os dados. Já esta pesquisa emprega 4 bases de dados distintas: Canvas, PhishTank, Alexa.com e DMOZ.

Diante do exposto, é válido destacar o que esta pesquisa tem de mais semelhante com os trabalhos apresentados neste capítulo são as 17 características Canvas. Entretanto, vale ressaltar que foram adicionadas mais 24 delas para aprofundar a ideia de detecção de scripts Canvas *fingerprinting*.

.

Capítulo 4

Detecção de Scripts Canvas *Fingerprinting*

Este capítulo detalha o método proposto para a detecção de scripts Canvas *fingerprinting* em páginas Web. Para tanto, detalha as características envolvidas na elaboração da proposta e a implementação dos mecanismos necessários para sua validação.

4.1 Método Proposto

O método proposto objetiva realizar a detecção de scripts Canvas *fingerprinting* em páginas Web baseando-se no modelo vetorial, a partir de um conjunto de características relevantes capazes de discriminar o mecanismo de Canvas *Fingerprinting*, extraídas por meio da análise estática do conteúdo da URL e do documento que compõe a página Web. Para tanto, emprega técnicas de Recuperação da Informação, introduzidas no Capítulo 2, para viabilizar a extração de características, a fim de realizar ao final o cálculo da similaridade e o ranqueamento dos scripts. Desta maneira, a extração de características, baseadas no código HTML, detecta e extrai os códigos em JavaScript, de forma a obter um conjunto de características relevantes na detecção e avaliação de scripts Canvas *Fingerprinting* em páginas Web.

Nessa perspectiva, a Figura 4.1 exemplifica o método proposto, dividido-o nas

seguintes etapas: coleta e extração de scripts, extração de características Canvas *fingerprinting*, o cálculo da similaridade e o ranqueamento dos scripts.

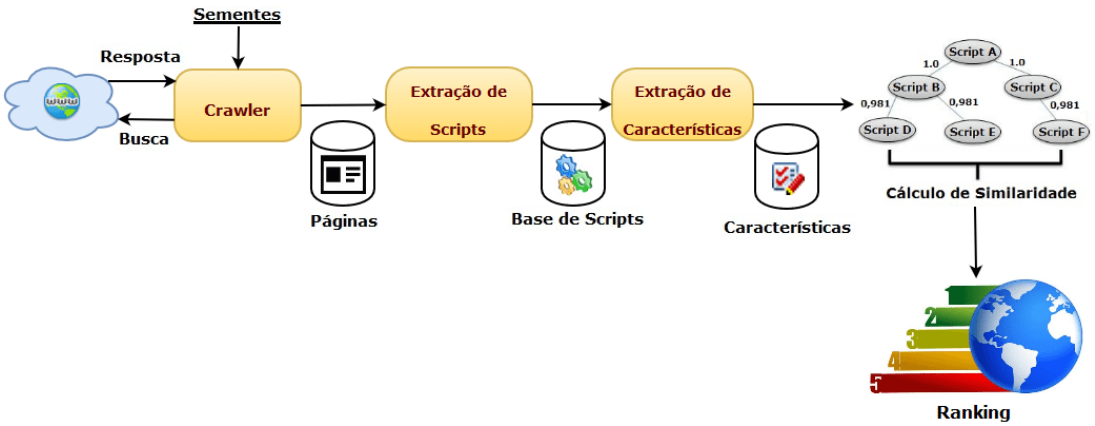

Figura 4.1: Método Proposto

4.1.1 Coleta e Extração de Scripts

A **Coleta e Extração de scripts** é realizada através de um ***Crawler***[1], que ao receber as sementes (URLs iniciais) faz o download somente dos scripts integrantes do código HTML das páginas. Os scripts coletados em Javascript são armazenados em uma base de dados. Em linhas gerais, o objetivo desta primeira fase é a construção da base de scripts.

É importante mencionar que as URL´s que possuam mais de um script terão seus scripts reunidos em um arquivo único que é armazenado com o nome da URL (como por exemplo, americanas ao invés do link inteiro), para que possa ocorrer a continuação das demais etapas do método proposto.

4.1.2 Extração de Características

A **Extração das Características** é realizada nos scripts coletados, através de um código extrator que detecta as características relacionadas ao Canvas *fingerprinting*. Ao final deste processo restarão somente os métodos e propriedades

[1]Crawler, também conhecido como Spider ou Bot, é um robô usado pelos buscadores para encontrar e indexar páginas de um site (GlobalAD)[31].

de Canvas *fingerpriting* existentes nos scripts analisados. Passo que irá a posteriori facilitar a realização do cálculo da similaridade. A lista de características extraíveis, incluindo descrição, será apresentada próxima seção. Esta fase tem a finalidade de construção da base de características Canvas *Fingerprinting*.

É importante mencionar que esta pesquisa não objetiva verificar códigos ofuscados, portanto, diversos scripts ao passarem pelo processo de extração de características obtiveram valoração em tamanho (0Kb), ou seja, os scripts estavam sem conteúdo, fato que inviabilizou utilizá-los.

4.1.3 Separação dos Scripts (Documentos e Consultas)

A **Separação dos scripts** é uma etapa manual, na qual após a coleta e extração dos scripts de uma base de links de Canvas, contendo 8.100 scripts, onde 8.000 foram retirados para compôr a base de dados Canvas (documentos d) desta pesquisa e os outros 100 scripts serviram para compôr as Consultas q. A escolha destas amostras de consulta foi totalmente aleatória, para que não influenciasse nos resultados desta pesquisa (as amostras variam de tamanho, tendo scripts com mais de 11.000 Kb até o mínimo de 100 Kb). Estes scripts que compõem as consultas q não se repetem em nenhuma das bases de dados utilizadas nesta pesquisa. Porém, eles servirão como scripts de comparação (contendo características Canvas *fingerprinting*) com os documentos (scripts das bases de dados).

4.1.4 Cálculo de Similaridade

O **Cálculo de Similaridade** é a etapa em que as propriedades e métodos Canvas *fingerprinting* dos scripts são apresentados como consulta ao índice por meio de um modelo de recuperação de informação (modelo vetorial).

Durante esse processo, o modelo vetorial é de suma importância, pois calcula o grau de similaridade do **documento (d)**, em relação à **consulta (q)** sob forma da correlação entre o **d** e **q**. Neste caso, cada **documento (d) da base Canvas *fingerprinting*** é representado como um vetor de termos e cada termo possui um valor associado que indica o grau de importância (peso) deste em um determinado **documento (d)**. De posse dos termos do vocabulário e das listas invertidas construídas é possível utilizar o modelo vetorial com vistas a

verificar o quão similar são os scripts existentes na base de dados em relação as 100 consultas. O cálculo do grau de similaridade, pautou-se na avaliação da função de *ranking*, do modelo Vetorial.

Como forma de ilustrar o cálculo de similaridade pelo método vetorial, a Tabela 4.1 apresenta o resultado de uma consulta (http://2016election.com) - conjunto de script de uma página Web - que é comparada com os 8.000 scripts da base Canvas (documentos), por meio da similaridade entre cossenos, através da equação 2.2 apresentada no Capítulo 2. É válido mencionar que o método vetorial realiza o cálculo da similaridade até para aqueles scripts que possuem pelo menos um termo igual ao da consulta. Entretanto, só são contabilizados e validados para este método, os valores que estejam entre o intervalo de similaridade de 0.1 até 1.0. Ou seja, os demais valores que estejam abaixo de 0.1 não são contabilizados.

Tabela 4.1: Tabela Demonstração do Cálculo de Similaridade

Tabela Demonstração do Cálculo de Similaridade			
Consulta	**Ordem no Ranking**	**Script Base Canvas**	**Nível de Similaridade**
http://2016election.com	1	Pistolsfiringbloy	0.9802
	2	Akdirahost	0.9539
	...	...	...
	324	Fieldgulls	0.4209
	...	...	...
	1724	Creativeplanetnetwork	0.1001
	...	...	...
	7992	Makespace	0.00368
	7993	Profilepic	0.00242

Na Tabela 4.1 são listados os scripts da base Canvas mais similares a consulta http://2016election.com e em seguida o valor do nível de similaridade desde o primeiro script (1) até o último (7.993), ou seja, de uma base com 8.000 scripts, somente 7 deles não tem nenhum nível de similaridade com a consulta em tela. Entretanto, como fora mencionado anteriormente, para o método vetorial somente são válidos os valores de similaridade entre 0.1 até 1.0. Portanto, só foram contabilizados como válidos até o script http://creativeplanetnetwork.com, número 1.724 do ranking, possuindo um nível de similaridade igual a 0.1001.

Esta listagem do primeiro até o último script apresentada anteriormente é denominada de ranqueamento, o qual será melhor elucidado na próxima subseção.

4.1.5 Ranking

O **Ranking** é a etapa em que, após realizado o cálculo da similaridade, os scripts são listados em um arquivo com o nome da consulta e o grau de similaridade de comparação com a base. O ranqueamento é exibido em ordem decrescente do nível de similaridade

Assim como na seção anterior, a Tabela 4.2 ilustra melhor o processo de ranqueamento, exibindo em sua segunda coluna, denominada de "Ordem de Script", o Ranking - a ordem em que os scripts estão ranqueados de acordo com o cálculo da similaridade realizado pelo método vetorial. Vale destacar que após a execução do método vetorial, responsável pelo cálculo da similaridade e o ranqueamento, o algoritmo *Knee Points* analisa os scripts listados no ranking e faz um corte daqueles que não são tão relevantes.

Tabela 4.2: Tabela Demonstração *Knee Points*

Tabela Demonstração *Knee Points*		
Consulta	**Ranking Inicial**	**Ranking após Algoritmo Joelho**
http://1057max.fm	2.743	349
http://2016election.com	1.724	324
http://androidtv.news	5.757	2.809
http://southendnewsnetwork.com	7.537	4.873
http://themamamaven.com	4.466	2.772

A Tabela 4.2 exibe apenas cinco (05) das 100 consultas, apresentando inicialmente o valor resultante de *matching* do método vetorial e, em seguida, após a aplicação do algoritmo do joelho, a lista de scripts que são semelhantes. Embora perceba-se uma queda drástica no número de scripts após a execução do algoritmo *Knee Points*, ainda seria basta difícil demonstrar os resultados, como visto na consulta `http://southendnewsnetwork.com`.

Ainda em relação a Tabela 4.2 é possível confirmar que a consulta `http://2016election.com` que fora apresentada na seção anterior na Tabela 4.1 obteve uma redução de 3.724 scripts ranqueados inicialmente pelo método vetorial e após a execução do algoritmo *Knee Points* obteve a redução para 324 scripts da base de dados Canvas ranqueados em comparação com esta consulta.

Assim, esta pesquisa, com base no ranking do método vetorial auxiliado pelo algoritmo joelho, apresenta os cinco primeiros scripts (TOP 5) para cada uma das 100 consultas, visto que seria incabível de modo textual listar cada um dos

scripts das bases de dados que foram ranqueados por consultas.

A Tabela 4.3 destaca o resultado dos TOP 5 scripts mais similares a consulta http://2016election.com, após o algoritmo *Knee Points*.

Tabela 4.3: Tabela Demonstração do Cálculo de Similaridade/Ranking

Tabela de Demonstração do Cálculo de Similaridade/Ranking			
Consulta	**Ordem no Ranking**	**Script Base Canvas**	**Nível de Similaridade**
http://2016election.com	1	Pistolsfiringblog	0.9802
	2	Akdirahost	0.9539
	3	Tap-repeatedly	0.9477
	4	Fdlreporter	0.9457
	5	Sheboyganpress	0.9457

A respeito das consultas, inicialmente fora pensado de modo empírico em fazer uso de 10% da quantidade de scripts de cada base de dados. Entretanto, após os experimentos, infelizmente tornou-se inviável a inclusão destes dados na dissertação, visto que para 250 consultas teria-se no mínimo 15 laudas só para apresentar uma tabela de similaridade de uma das bases que nem era a maior que utilizou-se, a qual possui 8.000 scripts. Em outras palavras, teriam-se 800 scripts de consultas, que resultariam em mais de 45 páginas para demonstrar uma tabela do nível de similaridade.

Outra questão importante são os TOP 5, que destaca os cinco primeiros scripts da base ranqueados para cada uma das 100 consultas. Vale informar que para um valor menor, três (TOP 3) por exemplo, quase não foi possível observar a variação de valores do ranking. Por outro lado, valores maiores (TOP 10 ou TOP 15, por exemplo) resultaram em uma tabela grande que também seria inviável de ser incluída nesta dissertação.

O próximo capítulo desta pesquisa apresenta os resultados experimentais obtidos a partir do emprego das características e do método proposto.

4.2 Detalhes de Implementação

De maneira prática, a implementação da proposta emprega:

- **Crawler**: foi desenvolvido um script em linguagem PHP (versão 5.5.15), principalmente por ser adequada para o desenvolvimento Web, capaz de realizar o download dos códigos em JavaScript das páginas.

- **Extração de Características**: foi desenvolvido um script em linguagem Phyton 3.4 para extrair as características, por meio de expressões regulares, executando-as nos códigos JavaScript. As expressões regulares são um excelente recurso em casamento de padrões e são amplamente usadas em conteúdo de texto, conforme menciona Sudkamp [32] e em detecção de anomalias como destaca Mayer [5]. Assim, ao final dessa extração restarão somente as propriedades e métodos Canvas *fingerprinting*, conforme mencionado na seção 4.1.2.

- **Cálculo de Similaridade**: um script em Python 3.4, o qual baseia-se na base de scripts Canvas *fingerprinting* e nas consultas, irá realizar uma comparação entre a base Canvas (a qual possui 8.000 scripts Canvas *fingerprinting*) e as consultas (as quais possuem 100 scripts Canvas *fingerprinting* extraídos via Crawler, de uma base de dados da Universidade de Princenton [2]).

- **Ranking**: o mesmo script em Python 3.4 mencionado na etapa anterior, o qual faz o cálculo da similaridade, após realizar o cálculo irá dispor em arquivos rotulados pelo nome das consultas (100) conforme o grau de similidade de comparação com a base (Canvas = 8.000 scripts, por exemplo), em formato de ranking do maior para o menor nível de similaridade (de 0.1 até 1.0).

4.3 Características Canvas *Fingerprinting*

As características estão compostas por propriedades e métodos Canvas *fingerprinting* totalizando 41. As Tabelas 4.4 e 4.5 destacam, respectivamente, as principais propriedades e métodos da API Canvas utilizadas para realizar o *fingerprinting*.

[2]https://webtransparency.cs.princeton.edu/webcensus/index.html

Tabela 4.4: Propriedades Canvas

Nome	**Descrição**
fillStyle	O fillStyle são conjuntos de propriedades que retornam a cor, gradiente ou padrão usada para preencher o desenho.
Canvas.font	Os conjuntos de propriedades do tipo font retorna as propriedades de fonte atuais para conteúdo de texto na tela.
textBaseline	Os conjuntos de propriedades TextBaseline retornam a linha de base do texto atual usado na elaboração do texto.
Canvas.width	Um inteiro positivo que reflete o atributo HTML largura do elemento, interpretados em pixels CSS. Quando o atributo não for especificado, ou se ele é definido como um valor inválido, como um negativo, o valor padrão de 300 é usado.
Canvas.height	Um inteiro positivo que reflete o atributo de altura HTML do elemento, interpretados em pixels CSS. Quando o atributo não for especificado, ou se ele é definido como um valor inválido, como um negativo, o valor padrão de 150 é usado.
strokeStyle	Os strokeStyle são conjuntos de propriedades que retornam a cor, gradiente ou padrão usado para golpes.
globalAlpha	São conjuntos de propriedades que retornam o valor alfa ou transparência atual do desenho.
lineWidth	Os conjuntos de propriedades LineWidth retornam a largura da linha atual, em pixels.
lineCap	Os lineCap são conjuntos de propriedades que retornam o estilo das tampas para uma linha.
lineJoin	Propriedade que define/retorna o tipo de canto criado, quando duas linhas se encontram.
miterLimit	Os miterLimit são conjuntos de propriedades que retornam o comprimento máximo de esquadria. Esse comprimento é a distância entre o canto interno e do canto externo, onde duas linhas se encontram.
shadowOffsetX	São conjuntos de propriedades que retornam a distância horizontal da sombra da forma.
shadowOffsetY	São conjuntos de propriedades que retornam a distância vertical da sombra da forma.
shadowBlur	Os conjuntos de propriedades shadowBlur retornam o nível de borrão para as sombras.
shadowColor	Os conjuntos de propriedades ShadowColor que retornam a cor a ser usada para sombras.
globalCompositeOperation	A propriedade CanvasRenderingContext2D.globalCompositeOperation da API Canvas 2D define o tipo de composição / operação para aplicar ao desenhar novas formas, cujo o tipo é uma string que identifica qual das operações de composição ou modo de mesclagem será utilizado.
textAlign	São conjuntos de propriedades que retorna o alinhamento atual para o conteúdo do texto, de acordo com o ponto de ancoragem.

Tabela 4.5: Métodos Canvas

Nome	Descrição
fillRect	O método fillRect desenha um retângulo "preenchido", cuja cor de preenchimento padrão é preta.
fillText	O método fillText desenha um texto preenchido na tela, cuja cor padrão do texto é preta.
lineTo	O método lineTo, adiciona um novo ponto e cria uma linha nesse ponto a partir do último ponto especificado na tela (este método não desenha a linha).
arcTo	Método que cria um arco/curva entre duas tangentes na tela.
beginPath	Método que inicia um caminho ou redefine o caminho atual.
clearRect	Método que limpa os pixels especificados dentro de um dado retângulo.
createImageData	Método CanvasRenderingContext2D.createImageData que pertence a API do Canvas 2D cria um novo objeto ImageData vazio com as dimensões especificadas. Todos os pixels do novo objeto são negros transparentes.
createPattern	Método que repete o elemento especificado na direção especificada.
createRadialGradient	Método que cria um objeto de gradiente radial / circular. Este gradiente pode ser usado para preencher retângulos, círculos, linhas, texto, etc.
measureText	Método que retorna um objeto o qual contém a largura do texto especificado, em pixels.
putImageData	Método CanvasRenderingContext2D.putImageData,pertencente a API do Canvas 2D pinta dados do objeto ImageData fornecido no bitmap. Se um retângulo sujo é fornecido, apenas os pixels desse retângulo são pintados.
quadractiveCurveTo	O método quadraticCurveTo adiciona um ponto ao caminho atual usando os pontos de controle especificados que representam uma curva quadrática de Bézier. A curva quadrática de Bézier requer dois pontos. O primeiro ponto é um ponto de controle que é usado no cálculo quadrático de Bézier e o segundo ponto é o ponto final da curva. O ponto de partida para a curva é o último ponto no caminho atual. Se um caminho não existir, usa os métodos beginPath () e moveTo () para definir um ponto de partida.
restore	O método CanvasRenderingContext2D.restore pertencente a Canvas 2D API restaura o estado da tela salva mais recentemente estalando a entrada superior na pilha de estado de desenho. Se não houver nenhum estado salvo, este método não faz nada.
rotate	O método CanvasRenderingContext2D.rotate pertencente a API do Canvas 2D adiciona uma rotação à matriz de transformação. O ângulo representa um ângulo de rotação no sentido horário e é expresso em radianos.
scale	O método CanvasRenderingContext2D.scale da API 2D Canvas adiciona uma transformação de escala para as unidades de tela por x horizontalmente e por y verticalmente. Por padrão, uma unidade na tela é exatamente um pixel. Se aplicarmos, por exemplo, um fator de escala de 0,5, a unidade resultante seria 0,5 pixels e assim as formas seriam desenhadas a meio tamanho. De forma semelhante, ajustar o fator de escala para 2,0 aumentaria o tamanho da unidade e uma unidade agora se tornaria dois pixels. Isso resulta em formas sendo desenhadas duas vezes maior.
setTransform	O método CanvasRenderingContext2D.setTransform pertencente a Canvas 2D API redefine (substitui) a transformação atual para a matriz de identidade e, em seguida, invoca uma transformação descrita pelos argumentos desse método.
strokeRect	O método strokeRect desenha um retângulo (sem preenchimento). Cuja cor padrão do traço é preta.
strokeText	O método strokeText, desenha texto (sem preenchimento) na tela. Cuja cor padrão do texto é preta.
getImageData	O método getImageData retorna um objeto ImageData que copia os dados de pixel para o retângulo especificado em uma lona.
toDataURL	Retorna um URL de dados que contém uma representação da imagem no formato especificado pelo tipo de parâmetro (o padrão é png). A imagem retornada está em uma resolução de 96 dpi.
getElementById('Canvas')	Retorna um objeto que fornece métodos e propriedades para desenhar e manipular imagens e gráficos em um elemento de tela em um documento. Este objeto inclui informações sobre cores, larguras de linha, fontes e outros parâmetros gráficos que podem ser desenhados em uma tela.
Canvas.getContext	Retorna um contexto de desenho na tela, ou nulo, se a identificação de contexto não é suportado. Um contexto de desenho permite desenhar na tela. Chamando getContext com "2d"retorna um objeto CanvasRenderingContext2D, enquanto que chamá-lo com "experimental-webgl"(ou "webgl") retorna um objeto WebGLRenderingContext. Neste contexto está disponível somente em navegadores que implementam WebGL.
createElement('Canvas')	Método criado para que sejam inclusos os elementos Canvas na página HTML.
getElementsByTagName-('Canvas')	Método que retorna uma coleção de todos os elementos no documento com o nome de tag especificado, como um objeto NodeList.

Capítulo 5

Experimentos e Resultados

Este capítulo descreve o protocolo experimental (ambiente e bases de dados) necessário para a realização da pesquisa descrita nesta dissertação, bem como os resultados alcançados.

5.1 Protocolo Experimental

Para avaliar o método proposto foram realizados vários experimentos, por meio de uma análise passiva nos scripts de cada uma das bases de dados, empregando todas as 41 características (propriedades e métodos Canvas *fingerprinting*) que constam na seção 4.3.

Contudo, antes de apresentar esses resultados é necessário descrever o ambiente, as bases de dados e o processo, que estão descritos em detalhes nas próximas subseções.

5.1.1 Ambiente

Os experimentos realizados foram executados em dois computadores. O primeiro é uma estação de trabalho Intel Core i7 de 2.7 Ghz, com 8 GB de memória RAM e disco SATA de 1 TB de armazenamento sob a plataforma Linux, distribuição Ubuntu 11.10. O segundo um notebook Intel Core i3 de 2.4 Ghz, com 4 GB de memória RAM e disco 250 GB de armazenamento com sistema operacional Windows 10.

5.1.2 Bases de Dados

Esta pesquisa baseia-se em quatro (4) bases de dados para realização dos experimentos, Canvas, Phishtank, DMOZ e Alexa, que serão explicitadas nas próximas subseções.

Optou-se por trabalhar não somente com as bases maliciosas, mas também com as bases não maliciosas, visando demonstrar que o Canvas *fingerprinting* está presente nos mais diversificados tipos de ambientes. E ainda, porque futuramente este trabalho pode ser implementado a um plug-in que informe aos usuários de websites se estão em perigo ou não, assim como ocorre nos antivírus por exemplo.

É válido mencionar que todas estas bases possuem tipos diversificados de sites, fato que permite destacar que o Canvas *fingerprinting* pode estar presente em qualquer um desses tipos de sites. Conforme ressalta Nikforakis [33], o *fingerprinting* está presente em vários tipos de categorias de sites como os de: compras, viagens, serviços na Internet, negócios/economia, entretenimento, encontros/namoro, Internet e computador, pornografia, maliciosos/spam.

A) Canvas

Esta base de dados é oriunda do trabalho de Englehardt et al. [14]. Originalmente, a base possui mais de 16.000 sites com Canvas, mas foi decidido utilizar somente sites com scripts de tamanhos iguais ou superiores a 100 Kb, como forma de garantir uma quantidade maior de informações para a realização da análise. Assim, a base foi composta inicialmente com 10.705 sites e após a realização do processo de RI, por meio do método vetorial, trabalhou-se somente com 8.100 scripts (sendo 8.000 para a Base Canvas e 100 para as consultas), visto que parte destes possuíam seus códigos ofuscados e como não é o foco desta pesquisa identificar este tipo de código, parte do que não houve resultado, fora eliminado.

Os scripts da base Canvas foram coletados nos meses de agosto a novembro de 2016.

B) Phishtank

A base Phishtank[1] pautou-se no site colaborativo de informações e dados relativos a *phishing* na Internet. O mesmo fornece uma API aberta para desenvolvedores e pesquisadores com vistas a integrar anti-*phishing* em suas aplicações, gratuitamente. Esta base é composta por mais de 20.000 links.

Para esta pesquisa, a amostra extraída desta base foi inicialmente de 4.366 scripts e após submetê-los ao processo de RI (método vetorial) trabalhou-se com 2.050 scripts. Esta diminuição na amostra ocorre pelo mesmo motivo descrito na base da alínea anterior. Os scripts da base phishtank foram coletados no interstício de novembro até dezembro de 2016.

Esta pesquisa procurou utilizar esta base de dados, por se tratar de uma API com links maliciosos de *phishing*, que conforme destaca Nunan [34], os ataques de *phishing* utilizam scripts que redirecionam usuários a uma página falsa, idêntica a original para obtenção de dados como: senhas, número do cartão de crédito e outros.

C) DMOZ

A base DMOZ[2] é um projeto de Diretório aberto, denominado Directory Mozilla (DMOZ), que conta com a colaboração de voluntários que editam e categorizam páginas da Internet. Esta pesquisa faz uso desta base de dados por se tratar de um diretório tido como uma base de sites considerados benignos. Esta base contém mais de 3.862,080 links das mais diversas categorias.

Para esta pesquisa, a amostra extraída desta base foi inicialmente de 1.577 scripts e após submetê-los ao processo de RI (método vetorial) trabalhou-se com 596 scripts. Esta diminuição na amostra ocorre pelo mesmo motivo descrito na base das alíneas anteriores. Os scripts da base DMOZ foram coletados no mês de Agosto de 2016;

[1] http://www.phishtank.com/
[2] http://DMOZ.org/

D) Alexa

A base Alexa[3] foi extraída da empresa Amazon que realiza um serviço na Web, o qual descrimina a quantidade de usuários que utilizam determinado site em certo momento e com base nesta informação a empresa realiza um *ranking* destes sites (Amazon) [35]. A empresa possui 10.000 sites cadastrados na categoria Brasil em sua base da dados.

Nesta pesquisa, a amostra extraída desta base de dados, fora inicialmente de 2.035 scripts e após submetê-los ao processo de RI (método vetorial) trabalhou-se com 1.478 scripts. Esta diminuição na amostra ocorre pelo mesmo motivo descrito na base da alínea anterior. Os scripts da base Alexa foram coletados no mês de setembro de 2016.

Para melhor elucidar os experimentos descritos nas próximas seções, é importante explanar sobre os cenários de avaliação, conforme destaque na seção 5.2.

5.2 Cenários de Avaliação

Esta seção destaca os três cenários de avaliação, os quais foram realizados em formato de experimentos: **Cenário 1** - Similaridade entre Scripts; **Cenário 2** - Consultas mais Relevantes; e **Cenário 3** - Top 5 (Nível de similaridade entre as bases de dados e as 100 consultas).

Desta maneira, para cada uma das quatro bases de dados foram realizados os experimentos nestes cenários, por meio do método vetorial e do algoritmo *knee points*.

Nesse sentido, as próximas subseções explicam de modo detalhado cada um dos cenários mencionados anteriormente.

5.2.1 Cenário 1 - Similaridade entre Scripts

Este experimento tem a finalidade de realizar uma contagem das características que aparecem nos resultados TOP 5 (resultado dos cinco primeiros scripts da base

[3]http://www.alexa.com/

que são semelhantes as 100 consultas) para cada consulta e verificar a quantidade de vezes que estes scripts aparecem no *ranking*, destacando os scripts das bases de dados que são mais similares aos 100 scripts das consultas.

Para exemplificar a ideia do experimento, considere que o site http://www.google.com.br esteja em uma das bases de dados analisadas. O método vetorial, ao fazer a comparação dos scripts desta base de dados com os scripts das 100 consultas, irá retornar que o script do site mencionado como exemplo, o "google" apareceu 10 vezes no ranking. A intenção é demonstrar, de maneira generalizada, a quantidade de vezes que os sites (scripts das bases de dados) apareceram no ranqueamento.

Para este cenário, devido ao volume de dados dos resultados, fez-se necessário o agrupamento de scripts das bases de dados de acordo com a quantidade de vezes que estes foram ranqueados, formando com isso, grupos para efeitos de demonstração.

5.2.2 Cenário 2 - Consultas mais Relevantes

Este experimento tem o intuito de demonstrar quais são as consultas que mais se repetem nas bases de dados.

É válido mencionar que o experimento só destaca os sites listados com base no método vetorial e no algoritmo *knee points*.

Para exemplificar, suponha que uma das 100 consultas desta pesquisa seja um script do site http://www.americanas.com.br e que após a execução do método vetorial comparando a base de dados com as consultas, retorne que o script mencionado anteriormente obtenha 700 scripts da base de dados que sejam relevantes para a consulta americanas. Ou seja, a consulta "americanas" possui 700 scripts da base de dados relevantes para ela.

É importante destacar que assim como fora mencionado no Capítulo 4, seção 4.1, subseções: 4.1.4 e 4.1.5, os valores considerados para o método vetorial são somente àqueles que estejam entre o intervalo de similaridade 0.1 até 1.0, respectivamente. Para efeito de demonstração, neste experimento as consultas foram agrupadas pela quantidade de scripts relevantes para cada uma das bases de dados.

5.2.3 Cenário 3 - Resultado Top 5 (Nível de Similaridade)

Este experimento visa demonstrar os cinco primeiros scripts da base de dados analisada no topo do ranking para cada uma das 100 consultas. Ao final, serão listadas para cada uma das 100 consultas, os TOP 5, ou seja, os cinco primeiros scripts da base de dados que está sendo analisada que são mais similares àquela consulta.

Tabela 5.1: Tabela Demonstração do Nível de Similaridade

Tabela Demonstração do Nível de Similaridade			
Consulta	**Ordem no Ranking**	**Script Base de Dados**	**Nível de Similaridade**
http://www.uol.com.br	1	Magazineluiza	1.0
	2	Hotmail	0.993
	3	Carrefour	0.968
	4	Mariza	0.842
	5	DB	0.837

Usando a Tabela 5.1 para melhor elucidar a ideia do experimento no cenário 3, suponha que o script do site `http://www.uol.com.br` seja uma das 100 consultas e que ao comparar com a base de dados obtenha-se como resultado do experimento os cinco primeiros scripts desta base, listados como mais semelhantes a esta consulta, destacando o seu respectivo nível de similaridade.

Para cada uma das bases de dados desta pesquisa foram calculados os níveis de similaridade, os quais encontram-se ao final desse documento do Apêndice E até o Apêndice H.

5.3 Resultados

Esta seção explicita os resultados alcançados na pesquisa acerca da comparação de cada um dos cenários mencionados anteriormente com os scripts das bases de dados mencionadas na seção 5.2. A seção subdivide-se de acordo com os cenários.

5.3.1 Resultados do Cenário 1

Antes de apresentar os resultados, é preciso esclarecer que, para efeito de demonstração, os scripts analisados foram agrupados pelo valor de vezes em que aparecem no topo do ranqueamento. Tal escolha se justifica pelo tamanho dos

resultados gerados em cada experimento, em formato de arquivo texto pela implementação, contendo mais de mil linhas.

Tomando como exemplo o resultado da base Canvas, foram geradas mais de 1.586 linhas, o que inviabilizaria a inserção destes dados de maneira isolada nessa dissertação.

Assim, as bases Canvas e Alexa, têm seus resultados organizados em seis grupos, a base Phishtank tem 13 grupos e a base DMOZ tem 12 grupos. A quantidade de grupos varia de acordo com as informações contidas nesse arquivo texto gerado para cada um dos experimentos com as bases de dados. Nos experimentos a seguir serão apresentadas tabelas contendo parte dos grupos definidos.

A) Canvas versus 100 Consultas

Este experimento irá confirmar se um script da base Canvas é mais semelhante a uma (ou mais) determinada(s) consulta(s), devido a quantidade de vezes que este apareceu no ranking. A Figura 5.1 demonstra a coleção de scripts da base Canvas, destacando a quantidade de vezes que apareceram no topo do ranqueamento.

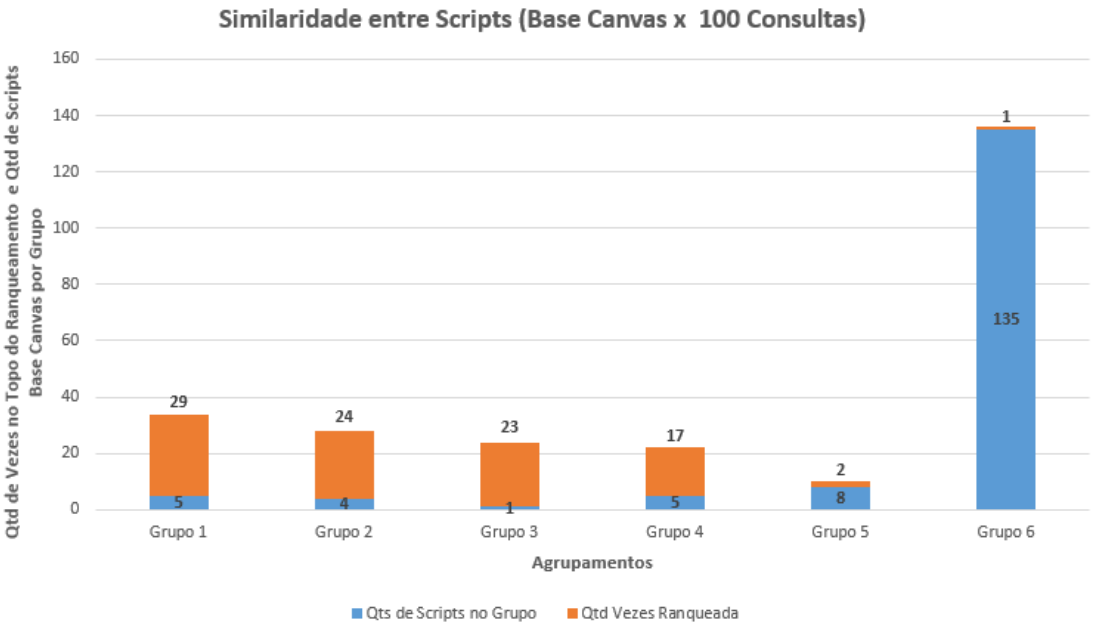

Figura 5.1: Similaridade entre Scripts (Base Canvas x 100 Consultas)

.

Percebe-se que os scripts dos grupos 1, 2 e 3 (Tabela 5.2) foram ranqueados mais de 23 vezes. Ou seja, os scripts destes grupos sempre fizeram parte do topo do ranking, comprovando o alto grau de similaridade destes scripts (base Canvas) com os 100 existentes nas consultas. Vale ressaltar que nesse experimento, os sites mais ranqueados nos grupos de 1, 2 e 3 são relacionados com: elaboração de vídeos, notícias, desenhos, filmes, decoração, guia técnico, consulados, entre outros.

Tabela 5.2: Grupos da Base Canvas

Similaridade entre Scripts (Base Canvas x 100 Consultas)			
Agrupamento de Scripts Base Canvas	**Scripts Base Canvas**	**Qtd Scripts no Grupo**	**Qtd Vezes Ranking**
Grupo 1	**guidingtech, ftlauderdalewebcam, preppyrunner, embassypages, nzhuntingandshooting**	**5**	**29**
Grupo 2	**sandraandwoo, filmhafizasi, dcclothesline, cleverlyinspired**	**4**	**24**
Grupo 3	**eastcoastcreativeblog**	**1**	**23**
...	...	...	...
Grupo 6	...	...	...

Esta constatação da diversidade de páginas contendo Canvas apenas comprova que não é possível distinguir a presença de Canvas *fingerprinting* sem analisar parte do conteúdo da página, visto que seria esperado que páginas com conteúdo gráfico mais visível seriam os principais alvos para Canvas, algo que não se comprovou como verdade.

B) Phishtank versus 100 Consultas

A Figura 5.2 ilustra a comparação da base Phishtank com as 100 consultas.

É fácil notar na Figura 5.2 que os grupos 1 e 2 tem apenas um script (`http://caraudioacapulco.com` e `http://sigarabirakmak.in`) em cada, sendo ranqueados 41 e 40 vezes, respectivamente. Já no grupo 3, os scripts `http://info-setting2016.twomini.com`, `http://ricardoeletro2.netai.net`, `http://www.infobel.com`, `http://maisponto.comxa.com` e `http://lagosstatenews.com` foram ranqueados 29 vezes. Por fim, no grupo 4, os scripts `http://rcacas.com`, `http://replacementroofingtx.com` e `http://davinciresidence.com.ar/br/` que foram ranqueados 24 vezes.

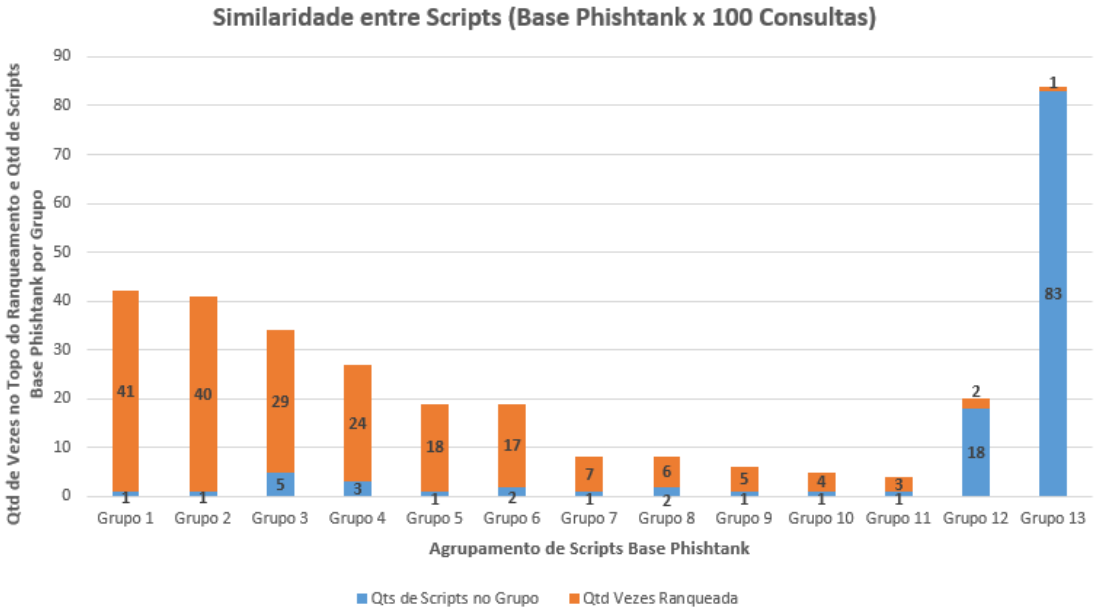

Figura 5.2: Similaridade entre Scripts (Base Phishtank x 100 Consultas)

.

Vale mencionar que entre os scripts mais ranqueados para esta base de dados encontram-se sites de: notícias, serviços, design, decoração, e-commerce e outros. Na Tabela 5.3 são listados os scripts da Base Phishtank dos quatro primeiros grupos mencionados na Figura 5.2. A Tabela completa com os grupos encontra-se no Apêndice B, sob a nomenclatura Tabela B.1.

Tabela 5.3: Grupos da Base Phishtank

Similaridade entre Scripts (Base Phishtank x 100 Consultas)			
Agrupamento de Scripts Base Phishtank	**Scripts Base Phishtank**	**Qtd Scripts no Grupo**	**Qtd Vezes Ranking**
Grupo 1	**caraudioacapulco**	**1**	**41**
Grupo 2	**sigarabirakmak**	**1**	**40**
Grupo 3	**info-setting2016, ricardoeletro2, infobel, maisponto, lagosstatenews**	**5**	**29**
Grupo 4	**replacementroofingtx, rcacas, davinciresidence**	**3**	**24**
...	...	...	...
Grupo 13	...	...	...

C) DMOZ versus 100 Consultas

A Figura 5.3 ilustra a comparação dos scripts da Base DMOZ com as 100 Consultas .

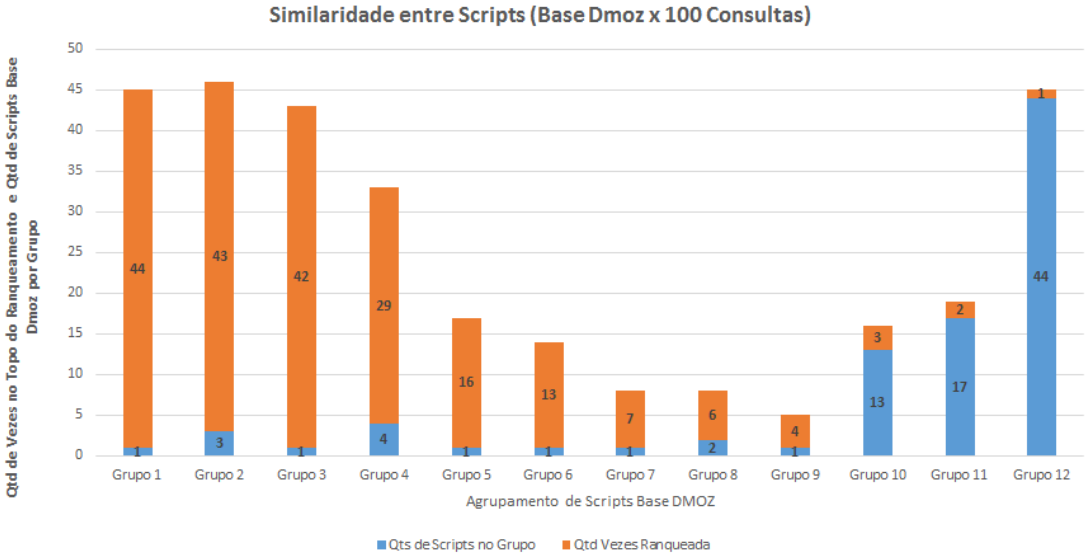

Figura 5.3: Similaridade entre Scripts (Base Dmoz x 100 Consultas)

.

Na Figura 5.3 percebe-se que o script http://www.mypet-memorial.net, grupo 1, foi ranqueado 44 vezes, enquanto os três scripts do grupo 2 (http://sonicyoga.com, http://www.modern-rocket.com e http://www.mdsafrica.net) foram ranqueados 43 vezes. Já o grupo 3, o qual possui o script http://panta-rhei.com, foi ranqueado 42 vezes e o grupo 4, o qual é composto por quatro scripts (http://bobthealien.co.uk, http://cottonclouds.com, http://shesmoke.blogspot.com.br e http://plum.tv), ranqueados 29 vezes. A Tabela 5.4 destaca exatamente os grupos mais similares.

É importante ressaltar que os sites que aparecem no topo do ranking são de categorias distintas: pet shop, religião/espiritualidade, oferta de serviços audiovisuais e serviço de teste de DNA, respectivamente. O Apêndice C trás a Tabela C.1 com todos os grupos e os resultados deste experimento.

Tabela 5.4: Grupos da Base Dmoz

Similaridade entre Scripts (Base Dmoz x 100 Consultas)			
Agrupamento de Scripts Base Dmoz	**Scripts Base Dmoz**	**Qtd Scripts no Grupo**	**Qtd Vezes Ranking**
Grupo 1	mypet-memorial	1	44
Grupo 2	sonicyoga, modern-rocket, mdsafrica	3	43
Grupo 3	panta-rhei	1	42
Grupo 4	bobthealien, cottonclouds, shesmoke, plum	4	29
...	...	...	...
Grupo 12	...	...	...

D) Alexa versus 100 Consultas

A Figura 5.4 apresenta a comparação dos scripts da Base Alexa com as 100 Consultas.

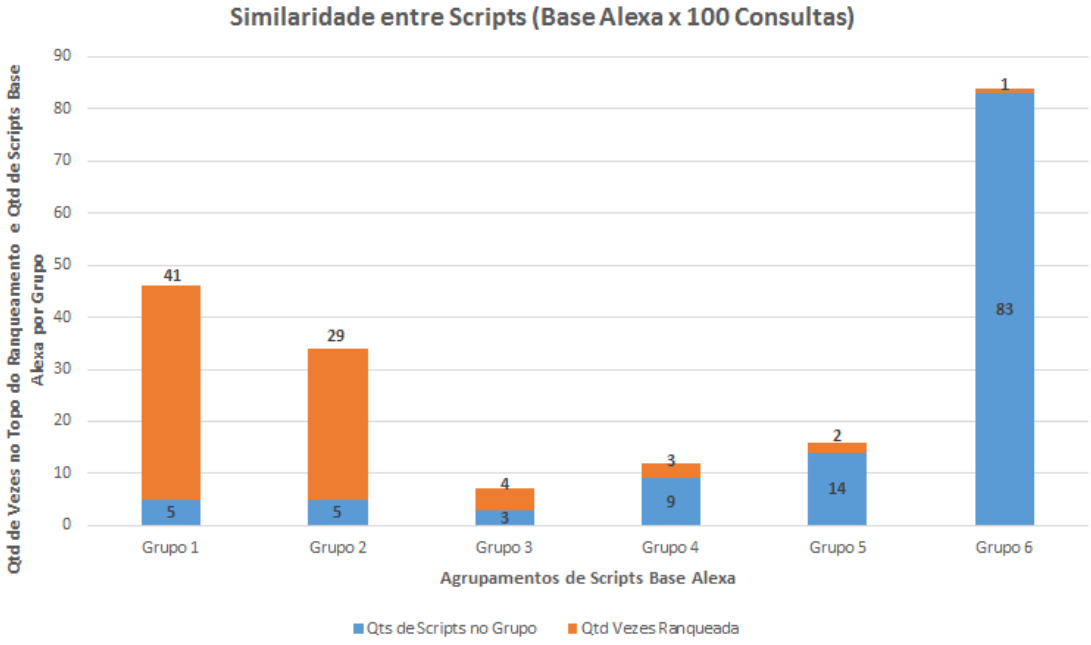

Figura 5.4: Similaridade entre Scripts (Base Alexa x 100 Consultas)

.

A Figura 5.4 destaca que apenas os grupos 1 e 2 são os mais relevantes. No grupo 1 existem scripts da base Alexa (http://www.agorams.com.br, http://aiesec.org.br, http://bigshopping.com.br e http://www.caadf.org.br) que foram ranqueados 41 vezes. No grupo 2 têm cinco scripts (http://www.

biomedicina-padrao.com.br, http://amofilmeshd.blogspot.jp, http://www.bussolaescolar.com.br, https://batepapo.uol.com.br e http://www.cidade-brasil.com.br) que são ranqueados 29 vezes. Em uma perspectiva mais detalhada, na Tabela 5.5, são apresentados os 6 grupos e os dois resultados com maiores valores. Dentre os scripts dos sites desta base de dados, são prevalentes sites de: conteúdo adulto, notícias, e-commerce, médico, filmes, bate papo, principais cidades do Brasil e escolar. O resultado completo encontra-se no Apêndice D, Tabela D.1.

Tabela 5.5: Grupos da Base Alexa

Similaridade entre Scripts (Base Alexa x 100 Consultas)			
Agrupamento de Scripts Base Alexa	**Scripts Base Alexa**	**Qtd Scripts no Grupo**	**Qtd Vezes Ranking**
Grupo 1	**adulto, caadf, agorams, aiesec, bigshopping**	**5**	**41**
Grupo 2	**biomedicinapadrao, amofilmeshd, batepapo, cidade-brasil, bussolaescolar**	**5**	**29**
...	...	...	...
Grupo 6	...	...	...

5.3.2 Discussão do Cenário 1

Os resultados do cenário em questão estão resumidos na Tabela 5.6.

Tabela 5.6: Discussão da Similaridade entre Scripts

Discussão Cenário 1 - Similaridade entre Scripts					
Base de Dados	**Total de Scripts na Base de Dados**	**Total de Grupos**	**Grupos com Maior Incidência no Ranking**	**Qtd Vezes Ranqueada**	**Total de Scripts nos Grupos de Maior Incidência no Ranking**
Canvas	8.000	6	1	29	5
			2	24	4
			3	23	1
Phishtank	2.050	13	1	41	1
			2	40	1
			3	29	5
			4	24	3
DMOZ	596	12	1	44	1
			2	43	3
			3	42	1
			4	29	4
Alexa	1.478	6	1	41	5
			2	29	5

É possível notar na Tabela 5.6 que para a base Canvas, dos seis grupos existentes, três aparecem com maior incidência no ranking, grupos 1, 2 e 3, ranqueados

29, 24 e 23 vezes, respectivamente. Em relação a base Phishtank, dos treze grupos, os quatro primeiros (totalizando 10 scripts) foram ranqueados 41, 40, 29 e 23 vezes, respectivamente. Na base DMOZ, com 12 grupos, os 4 primeiros tiveram maior prevalência. Por fim, na base Alexa, apenas 2 grupos dos 6 grupos se destacaram. Embora as bases Alexa e DMOZ tenham obtido resultados similares, com uma maior prevalência para primeira, percebe-se que os valores de ranqueamento são altos para a base DMOZ, que é benígna, ou seja, é de fato um resultado não esperado. Já na base Alexa, esperava-se um valor alto, uma vez que esta contém sites do Brasil dos tipos (conteúdo adulto, redes sociais, compras e etc.), conforme afirma Nikiforakis [33], que destaca que estes tipos de sites fazem uso da técnica de *fingerprinting*.

A explicação para a base benígna ter obtido um resultado de similaridade tão alto em comparação com as consultas tem a ver com o uso das caraterísticas de *fingerprinting* estarem presentes nos mais diversificados tipos de sites. Alguns apenas para ajustar conteúdos e em casos mais drásticos também capturar dados dos usuários que o utilizam estes sites.

Diante dos dados expostos na Tabela 5.6, na coluna **Total de Scripts nos Grupos de Maior Incidência no Ranking**, é possível notar que há uma coincidência na soma de scripts ranqueados por base, pois as bases: Canvas, Phishtank e Alexa possuem um total de dez (10) scripts, diferindo somente da base DMOZ que totaliza em 9 scripts. Uma explicação para esse fato pode residir na quantidade de consultas, já que a maioria das bases de dados possuem em média 1.300 scripts que se relacionam com 100 consultas **(fixas)**, enquanto a base Canvas com 8.000 scripts (maior de todas) também relaciona-se com estes 100 scripts da coleção de consultas.

Assim, com a finalidade de validar esse questionamento, fora realizado um experimento com a base Canvas utilizando 770 scripts de consulta. Este resultado, Tabela 5.7, prova que o tamanho das amostras de consulta tem influência no ranqueamento.

Tabela 5.7: Similaridade entre Scripts (Base Canvas x 770 Consultas)

Similaridade entre Scripts (Base Canvas x 770 Consultas)			
Agrupamento de Scripts Base Canvas	**Script Base Canvas**	**Qtd Scripts no Grupo**	**Qtd Vezes Ranqueada**
Grupo 1	**dcclothesline, filmhafizasi, cleverlyinspired, sandraandwoo**	**4**	**41**
Grupo 2	**eastcoastcreativeblog**	**1**	**32**
Grupo 3	**guidingtech, preppyrunner, embassypages, ftlauderdalewebcam, nzhuntingandshooting**	**5**	**30**
Grupo 4	**dykn**	**1**	**23**
Grupo 5	**teknotrik, whoneedsmaps, thefuturebuzz, techydroid**	**4**	**22**
...	...	...	...
Grupo 21	...	812	1

5.3.3 Resultados do Cenário 2

Para esse experimento, como forma de melhorar a demonstração dos resultados, agrupou-se as 100 consultas pelos valores da quantidade de scripts das bases de dados que são relevantes por consulta. Por exemplo, se três consultas forem relevante para 2.000 scripts da base Phishtank, estes são agrupados em um único grupo de consultas.

É válido destacar que os gráficos aqui expostos não tem finalidade estatística, visto que os valores dos resultados foram agrupados para efeito de demonstração.

A seguir nas alíneas de A até D, são apresentados os resultados dos experimentos realizados para este cenário.

A) Canvas versus 100 Consultas

A Figura 5.5 ilustra a quantidade de sites da base Canvas que são relevantes para cada uma das 100 consultas.

Vale destacar que só são listados os sites, com base no método vetorial e no *Knee points*.

A Figura 5.5 demonstra que a consulta 1, http://southendnews-network.com, apresenta um percentual de relevância de 61% em relação aos scripts da base de dados Canvas, os quais são representados por 4.873 scripts da base em questão que são relevantes para as consultas deste grupo. Também demonstra que, em média, os scripts são pelo menos 20% relevantes para as consultas dessa faixa de percentual, como as consultas 14 (http://news24eg.com) e 15 (http://news24zim.com) e finaliza destacando que mesmo no menor valor percentual,

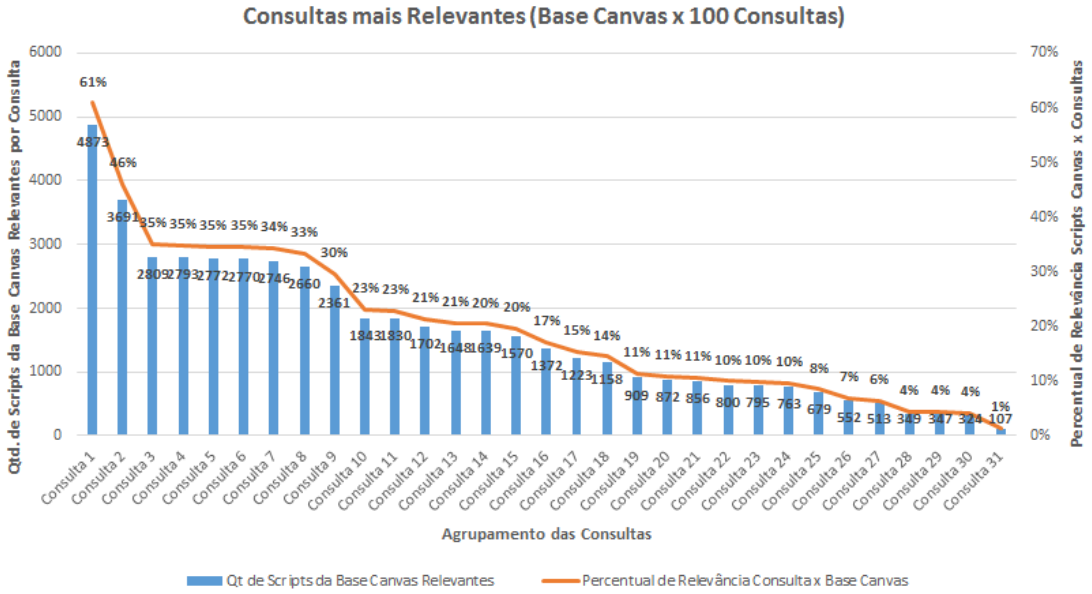

Figura 5.5: Consultas mais Relevantes (Base Canvas x 100 Consultas)

.

representado pela consulta 31 (http://najducokoliv.cz), têm-se pelo menos uma quantidade superior a 100 scripts da base Canvas que são relevantes para aquele grupo de consultas.

Para melhor ilustrar a ideia exposta na Figura 5.5, a Tabela A.2, Apêndice A, permite visualizar todos os scripts da base de dados em questão que foram ranqueados neste experimento.

B) Phishtank versus 100 Consultas

A Figura 5.6 ilustra a quantidade de scripts da base Phishtank que são relevantes para cada uma das 100 consultas, onde se destaca a consulta 1, http://south-endnewsnetwork.com, com 48% de relevância em relação a base de dados, tendo 975 scripts da base Phishtank que são relevantes para a consulta mencionada.

Em seguida, o script da consulta 2 (http://epicobottles.de) apresenta um percentual de relevância de 41% em relação a base Phishtank, representada por 843 scripts nesta base de dados. Um outro fato importante em pauta no final do gráfico é que para o menor valor percentual, representado pela consulta

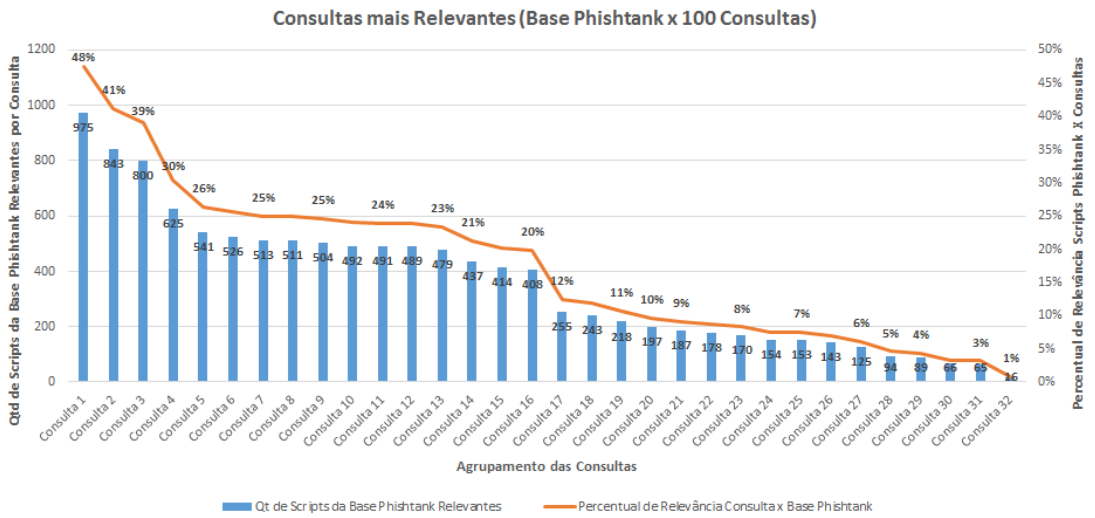

Figura 5.6: Consultas mais Relevantes (Base Phishtank x 100 Consultas)

.

32 (http://leisecamarica.com.br), têm-se ao menos uma quantidade de 16 scripts da base Phishtank que são relevantes para esta consulta.

O resultado completo deste experimento está em destaque no Apêndice B, Tabela B.2.

C) DMOZ versus 100 Consultas

A Figura 5.7 ilustra a quantidade de scripts da base Dmoz que são relevantes para cada uma das 100 consultas, onde é possível vislumbrar que a consulta 1, http://southendnewsnetwork.com, é 57% relevante em relação a base de dados, tendo 339 scripts da base Dmoz que são relevantes para esta consulta, seguido do script da consulta 2 http://quirkychrissy.com, a qual apresenta um percentual de relevância de 43% em relação a base Dmoz, representada por 258 scripts nesta base de dados. Ainda é possível visualizar neste gráfico, que a consulta 29, representada pelo script http://opovonews.com.br, têm o quantitativo de 15 scripts da base Dmoz que são relevantes para esta consulta.

O resultado completo deste experimento está em destaque no Apêndice C, na Tabela C.2.

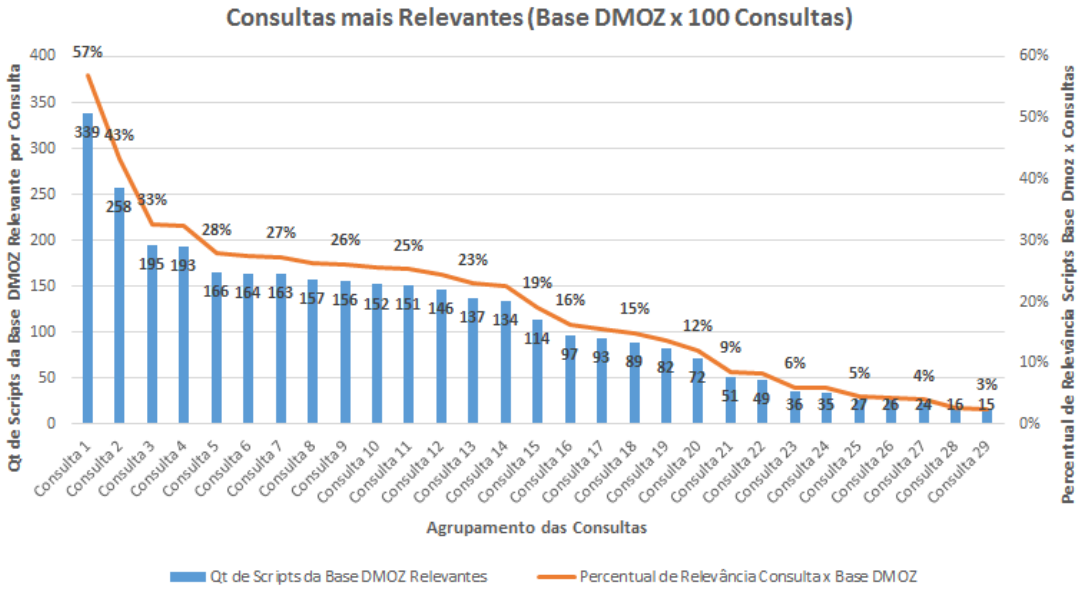

Figura 5.7: Consultas mais Relevantes (Base Dmoz x 100 Consultas)

.

D) Alexa versus 100 Consultas

A Figura 5.8 ilustra a quantidade de scripts da base Alexa que são relevantes para cada uma das 100 consultas, onde é possível vislumbrar que a consulta 1, http://quirkychrissy.com, é 42% relevante em relação a base de dados, tendo 620 scripts da base Alexa que são relevantes para esta consulta, seguido do script da consulta 2 (http://naivecookcooks.com), a qual apresenta um percentual de relevância de 40% em relação a base Alexa, representada por 595 scripts nesta base de dados. O gráfico ainda destaca que a consulta 30, representada pelo script http://leisecamarica.com.br, têm o quantitativo de 65 scripts da base Alexa que tem um percentual de 4% de relevância para esta consulta.

O resultado completo deste experimento está em destaque no Apêndice D, na Tabela D.2.

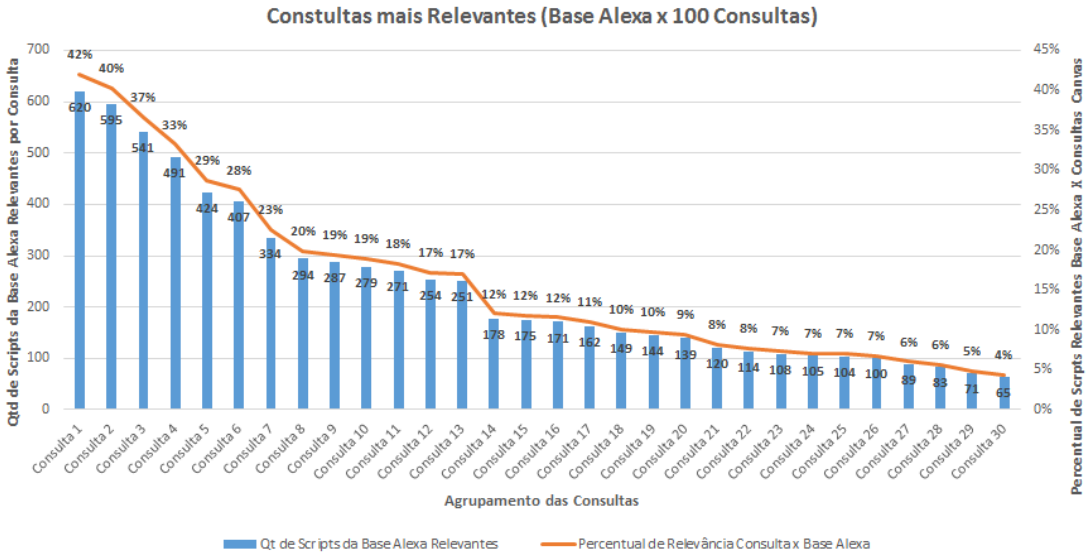

Figura 5.8: Consultas mais Relevantes (Base Alexa x 100 Consultas)

.

5.3.4 Discussão do Cenário 2

Os resultados do cenário 2 estão resumidos na Tabela 5.8, a qual demonstra as bases de dados, o total de scripts, a quantidade de consultas agrupadas, a quantidade de consultas relevantes, as URLs das consultas, as quantidades de scripts relevantes e o percentual de relevância.

Tabela 5.8: Tabela Discussão Cenário 2 - Consultas mais Relevantes da Base x 100 Consultas

Discussão Cenário 2 - Consultas mais Relevantes da Base x 100 Consultas						
Base de Dados	**Total de Scripts na Base de Dados**	**Qtd de Consultas Agrupadas**	**Qtd de Consultas Relevantes**	**Consulta**	**Qtd Scripts Relevantes por Base de Dados**	**Percentual de Relevância**
Canvas	8.000	31	2	southendnewsnetwork	4.873	61%
				epicobottles	3.691	46%
Phishtank	2.050	32	2	southendnewsnetwork	975	48%
				epicobottles	843	41%
DMOZ	596	29	2	southendnewsnetwork	339	57%
				quirkychrissy	258	43%
Alexa	1.478	30	2	quirkychrissy	620	42%
				naivecookcooks	595	40%

Percebe-se nos valores da tabela que a base Canvas possui duas consultas (`http://southendnewsnetwork.com` e `http://epicobottles.de`), as quais obtiveram um percentual de relevância de 61% e 46% respectivamente, sendo a

primeira relacionada 4.873 scripts da base e a segunda um total de 3.691 scripts. Na base Phishtank, as mesmas consultas são as mais relevantes, tendo percentuais de 48% (para 975 scripts) e 41% (para 843 scripts). Já para a base DMOZ, o script da consulta `http://southendnewsnetwork.com` também foi o mais relevante, obtendo um percentual de 57% com 339 scripts da base, a qual possui um total de 593 scripts. Tal fato gera bastante curiosidade, visto que esta base é tida como benigna. O outro script, da consulta `http://quirkychrissy.com`, apresenta um nível percentual de relevância de 43%. Por fim, na base Alexa, o script da consulta `http://quirkychrissy.com` tem resultado bastante similar ao da base de dados DMOZ, com 42% de relevância. E a segunda consulta, script `naivecookcooks`, tem relevância de 40%.

O fato real que faz estas consultas serem tão relevantes para cada uma das bases de dados mencionadas se deve à semelhança entre as características prevalentes entre elas, já que todas possuem em comum características como *fillText*, *textBaseline*, *toDataURL*, *clearRect*, *getImageData*. Para a consulta `http://southendnewsnetwork.com`, a diferença é que esta possui três características a mais: *scale*, *textAlign*, *restore*.

Já a consulta `http://epicobottles.de` têm apenas *scale* como característica adicional. Para a consulta `http://quirkychrissy.com` existe a mais as características *rotate*, *restore* e, por fim, a consulta `http://naivecookcooks.com` têm somente a característica *restore* assim como na consulta anterior.

Com relação a relevância das características mencionadas anteriormente, em sua maioria e de modo geral são características que transformam texto em linhas, retornam a linha de base do texto inicial, limpam os pixels especificados dentro de um retângulo e transformam o texto em uma imagem em base 64, igualmente a ideia mencionada no Capítulo 2 Seção 2.2 que trata do HTML Canvas. Por outro lado, em um estudo anterior a este, Saraiva [30] destaca que as características *getImageData* e *toDataURL* são classificadas com de alta periculosidade.

Já para as características encontradas além das supracitadas, há a que destaca o alinhamento atual do conteúdo do texto, a que faz a transformação de escala em tamanhos horizontal/vertical na tela, e àquela que restaura o estado da tela salva mais recentemente. Deste modo, é possível recuperar conteúdo e último estado apresentados na tela do usuário. Para maiores informações sobre

as funcionalidade de cada uma das características mencionadas, vide seções 4.3 em 4.5 no Capítulo 4.

Diante dos fatos, pode-se afirmar que a consulta (http://southendnews-network.com) é bastante relevante para quase todas as bases. O interessante é que trata-se de um site de notícias de cunho mundial, simples e sem design arrojado. Ou seja, não causa suspeitas que realiza o Canvas *fingerprinting*, por não conter recursos multimídia atrativos.

5.3.5 Resultados do Cenário 3

Para este cenário, como forma de melhor ilustrar a ideia do experimento, foi criada uma legenda de cores de maneira empírica, com base na classificação da ISO 27.000 [36], do grupo de segurança da informação, onde a cor Vermelha representa um nível de risco altíssimo (90% à 100%), a cor Amarela representa um nível de risco médio (60% à 79%) e a cor Verde representa um nível de risco baixo (0% à 59%).

Neste trabalho foi adicionada a cor Laranja, a qual representa um nível de risco alto, em torno 80% até 89%. Esta classificação serve para dar ideia dos níveis de similaridade, fazendo um intervalo percentual diferenciado. Assim, fica mais nítido o intervalo de 90% até 100% e as demais classificações.

Vale destacar que nas imagens referentes aos resultados deste cenário, os retângulos coloridos com as cores supracitadas representam as consultas e os círculos brancos representam os documentos das bases de dados. Para especificar a cor dos retângulos, foram analisados os TOP 5 scripts mais similares às consultas e com base na frequência dos valores de similaridade adotou-se a cor. Por exemplo, na Tabela E.5 do Apêndice E apresenta a consulta http://guesstheemoji-answers.com que destaca um nível de similaridade com cinco scripts da base Canvas: blooshsports, tohapi, date, freeonlinephotoeditor e fastsocialfollower, os quais possuem um nível de similaridade de: 0.936, 0.921, 0.891, 0.884 e 0.868, respectivamente. O valor para a coloração da consulta que irá prevalecer é o da faixa de 80%, visto que a maioria dos níveis de similaridade são prevalentes nesse intervalo, portanto, a cor a ser pintada é a laranja.

A) Canvas versus 100 Consultas

O intuito deste resultado é demonstrar os 5 (cinco) primeiros scripts da Base de dados Canvas no topo do ranking para cada uma das consultas. A Figura 5.9 destaca o grafo de similaridade completo contendo a relação das 100 consultas Canvas com a Base de dados Canvas após a execução do método vetorial e do algoritmo *Knee Points*, representado pelos Top 5, cinco scripts da base Canvas, mais similares as consultas.

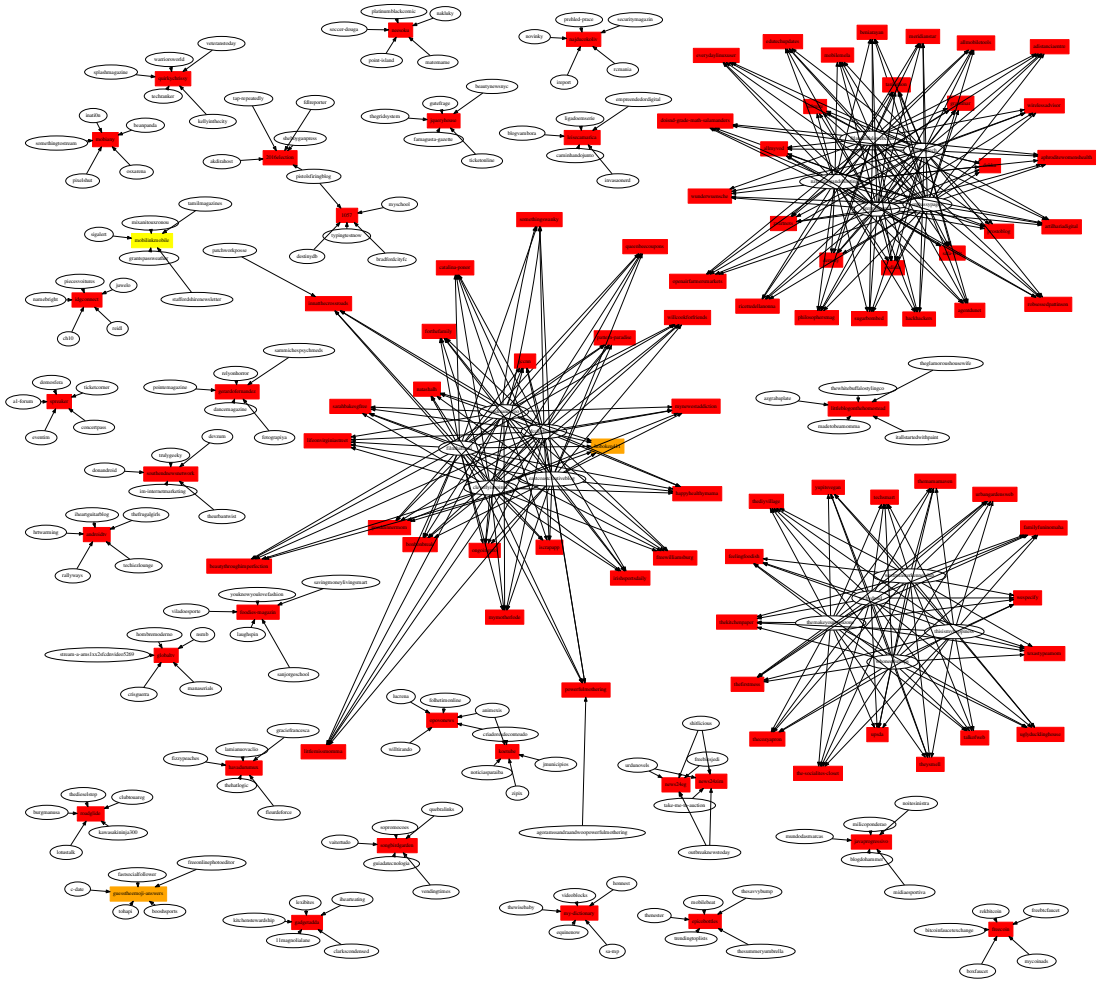

Figura 5.9: Grafo do Nível de Similaridade Base Canvas x 100 Consultas

.

Por outro ângulo, a Figura 5.10 demonstra uma parte do grafo de similaridade, na qual é possível visualizar 17 consultas que relacionam-se com apenas 5 (cinco)

scripts da base de dados Canvas, apresentando um nível de similaridade que varia de 90% até 100%.

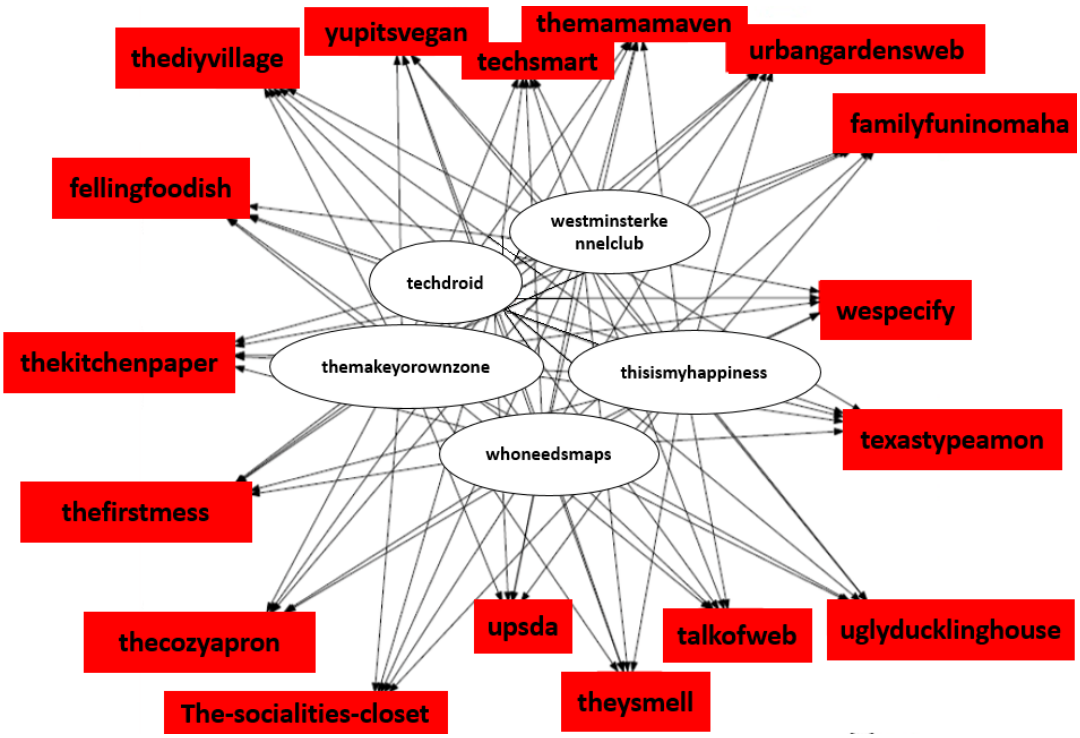

Figura 5.10: Parte do Grafo com o Nível de Similaridade Base Canvas x 100 Consultas

.

É válido frisar que tanto no grafo completo, Figura 5.9, quanto no grafo apresentado na Figura 5.10, a maioria das relações tem este mesmo grau de similaridade, o que deixa evidência de que a base de dados analisada é bastante similar as 100 consultas.

Para comprovar o nível de similaridade, o gráfico representado na Figura 5.11 destaca que a Base Canvas têm 97% de similaridade (variando de 90% até 100%) com as 100 consultas Canvas selecionadas aleatoriamente para realização deste experimento. Os demais valores expostos na imagem ressaltam 2% para o nível alto (variando de 80% até 89%), seguida de 1% para o nível médio (variando de 60% até 79%) e 0% para o nível baixo.

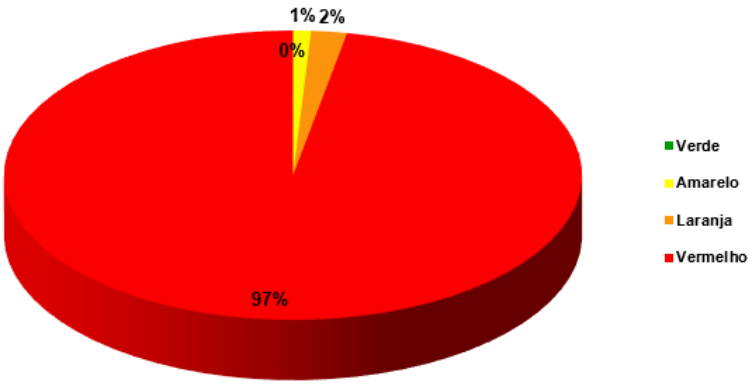

Figura 5.11: Gráfico do Nível de Similaridade Base Canvas x 100 Consultas

.

B) Phishtank versus 100 Consultas

A Figura 5.12 destaca o Grafo de similaridade completo contendo a relação das 100 consultas Canvas com a base de dados Phishtank após a execução do método vetorial e do algoritmo *Knee Points*, representado pelos Top 5, que são os 5 scripts da base phishtank, mais similares as 100 consultas. É importante mencionar que a maioria das consultas tem um alto nível de similaridade, visto que a prevalências da cor vermelha que representa um valor entre 90% a 100% de similaridade comparação entre base e consultas.

Já a Figura 5.13 demonstra a relação de similaridade entre três consultas com três scripts da base Phishtank, os quais, além de estabelecer uma relação entre as consultas mencionadas anteriormente, ainda têm um altíssimo nível de similaridade em torno de 90% a 100%.

Na imagem do grafo completo, a maioria das relações tem este mesmo grau de similaridade, o que deixa nítido que a base de dados analisada é bastante similar as 100 consultas. Para explicitar o nível de similaridade, o Gráfico representado na Figura 5.14, destaca que a Base Phishtank têm 87% de Similaridade de altíssimo nível (variando de 90% até 100%) com as 100 consultas selecionadas aleatoriamente para realização deste experimento, os outros percentuais somados chegam a 13% pontos percentuais.

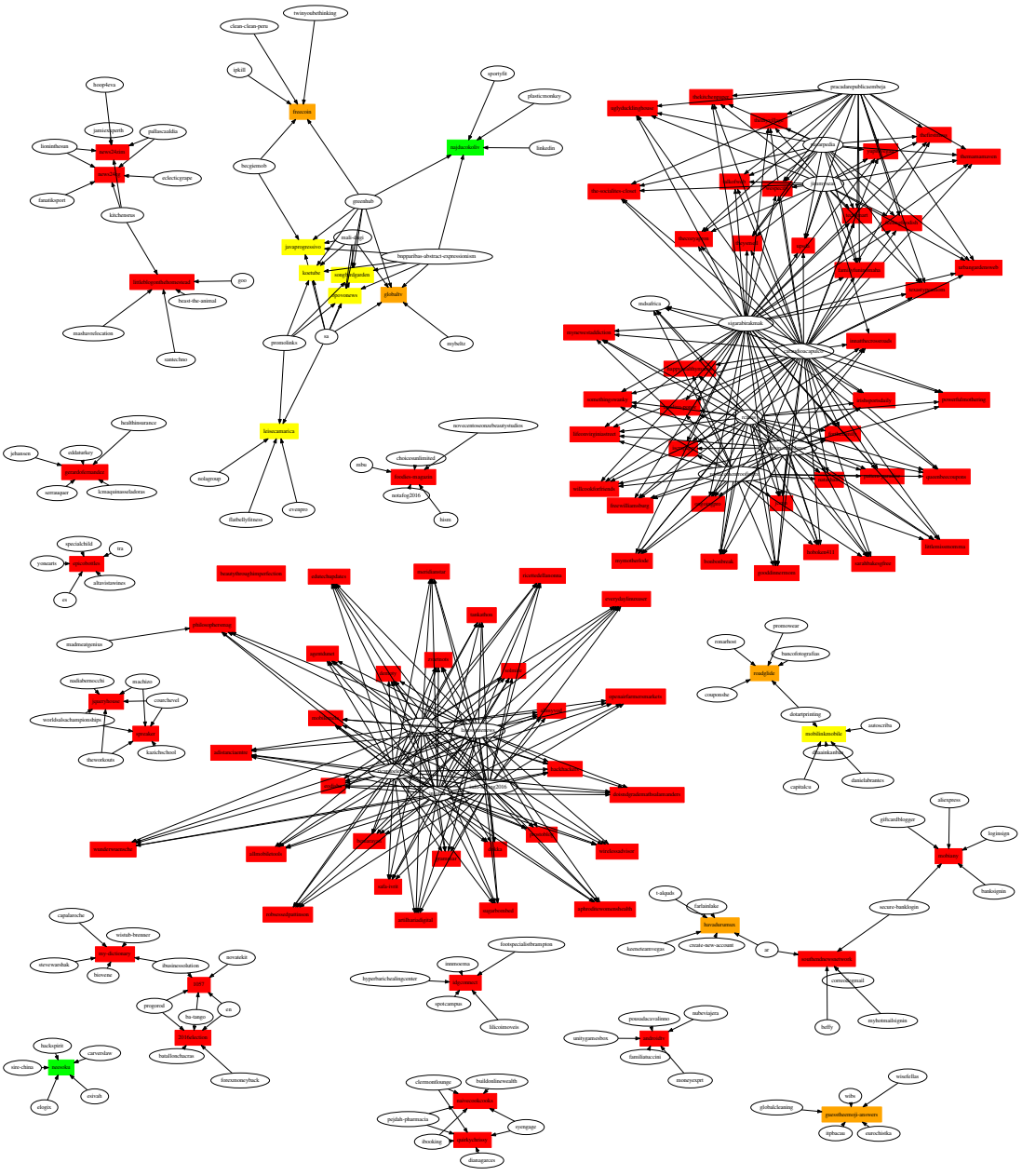

Figura 5.12: Grafo do Nível de Similaridade Base Phishtank x 100 Consultas

.

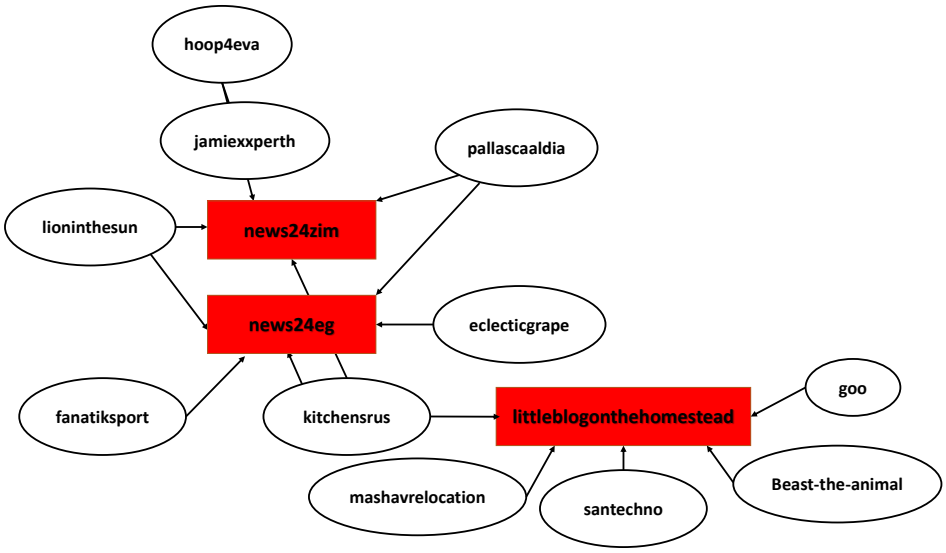

Figura 5.13: Parte do Grafo com o Resultado do Nível de Similaridade Base Phishtank x 100 Consultas

.

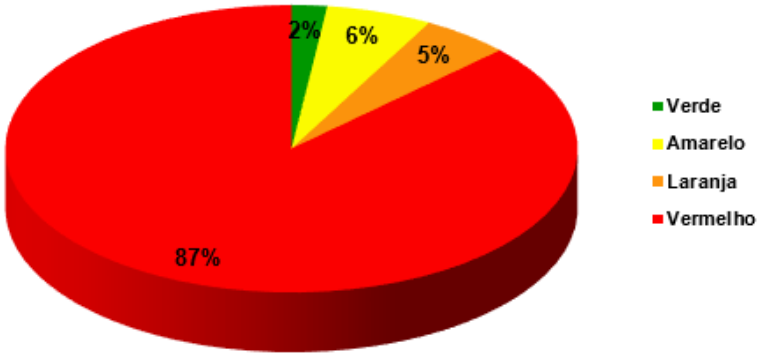

Figura 5.14: Gráfico do Nível de Similaridade Base Phishtank x 100 Consultas

.

C) DMOZ versus 100 Consultas

A Figura 5.15 apresenta o Grafo de similaridade completo contendo a relação das 100 consultas com a Base de dados Dmoz após a execução do método vetorial e do algoritmo *Knee Points*, representado pelos Top 5, ou seja, os 5 scripts da base Dmoz, mais similares as consultas. É válido destacar que a maioria das consultas tem um altíssimo nível de similaridade, visto que na imagem é prevalente a cor vermelha que representa um valor entre 90% a 100% de similaridade na comparação entre base de dados e as 100 consultas.

Sob um outro ângulo de visualização, a Figura 5.16 demonstra que há correlação entre quatro (4) scripts das consultas Canvas, os quais relacionam-se com os scripts da base Dmoz. Os scripts citados são relacionados a sites de treinamento de gestão de vendas e negociações, depoimento de negligência médica, gastronomia e turismo, respectivamente.

Em relação ao percentual do nível de similaridade, o experimento demonstra que a base Dmoz é 87% similar as 100 consultas Canvas, seguido de 4% do nível que varia de 80% a 89% de similaridade; e ainda 8% do nível que varia 60% até 79%; e 1% do nível que varia de 0% até 59%. Ou seja, esta base de dados destaca um altíssimo nível de similaridade, pois o valor apresentado no gráfico, representa um nível entre 90% a 100% de similaridade.

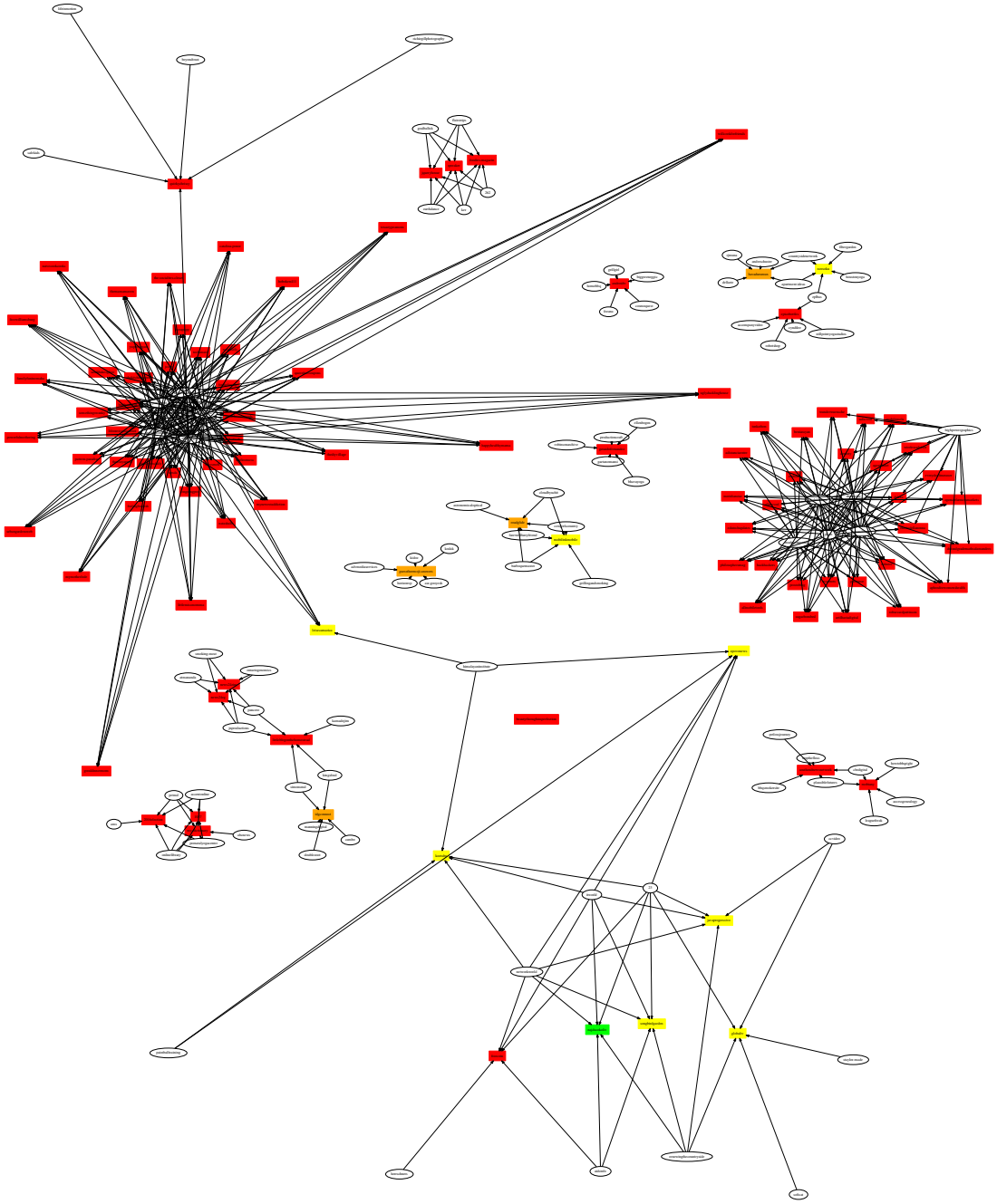

Figura 5.15: Grafo do Nível de Similaridade Base Dmoz x 100 Consultas

.

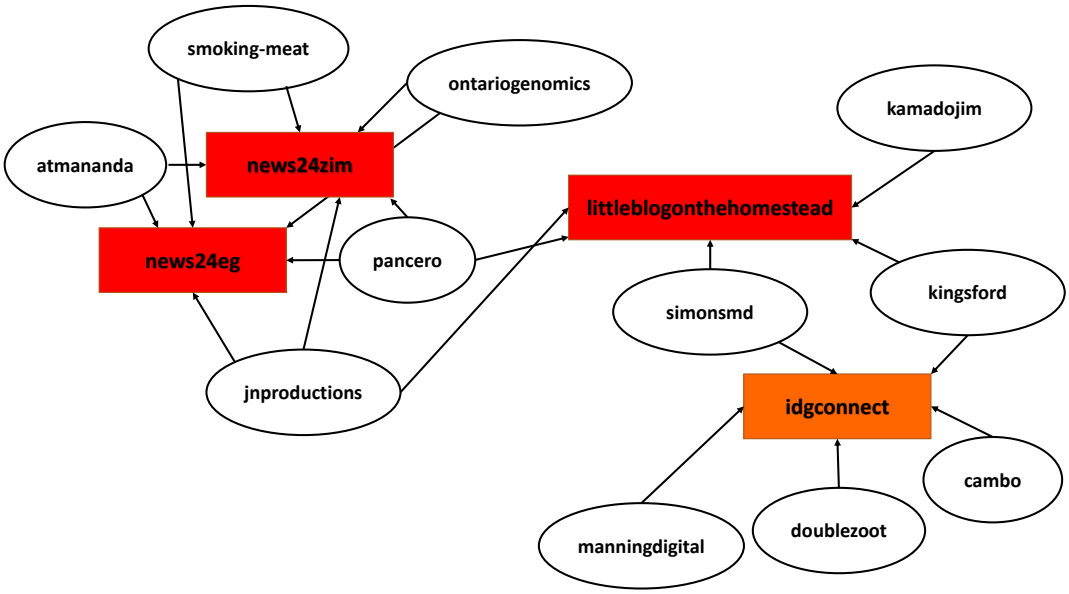

Figura 5.16: Parte do Grafo com o Resultado do Nível de Similaridade Base Dmoz x 100 Consultas

.

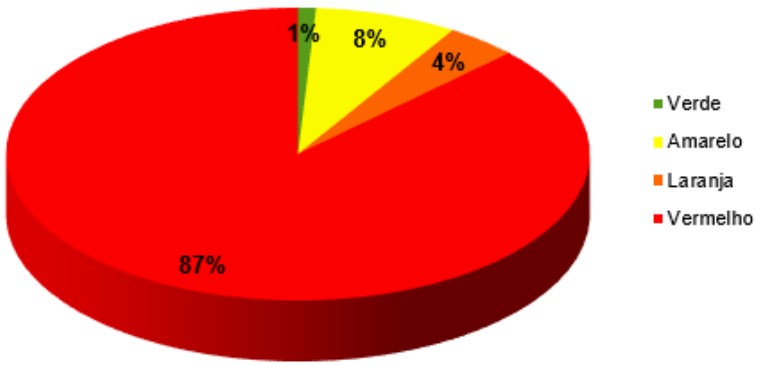

Figura 5.17: Gráfico do Nível de Similaridade Base Dmoz x 100 Consultas

.

D) Alexa versus 100 Consultas

A Figura 5.18 destaca o Grafo de similaridade completo contendo a relação das 100 consultas com a Base de dados Alexa após a execução do método vetorial e do algoritmo *Knee Points*, representado pelos Top 5, ou seja, os 5 scripts da base Alexa, mais similares as 100 consultas.

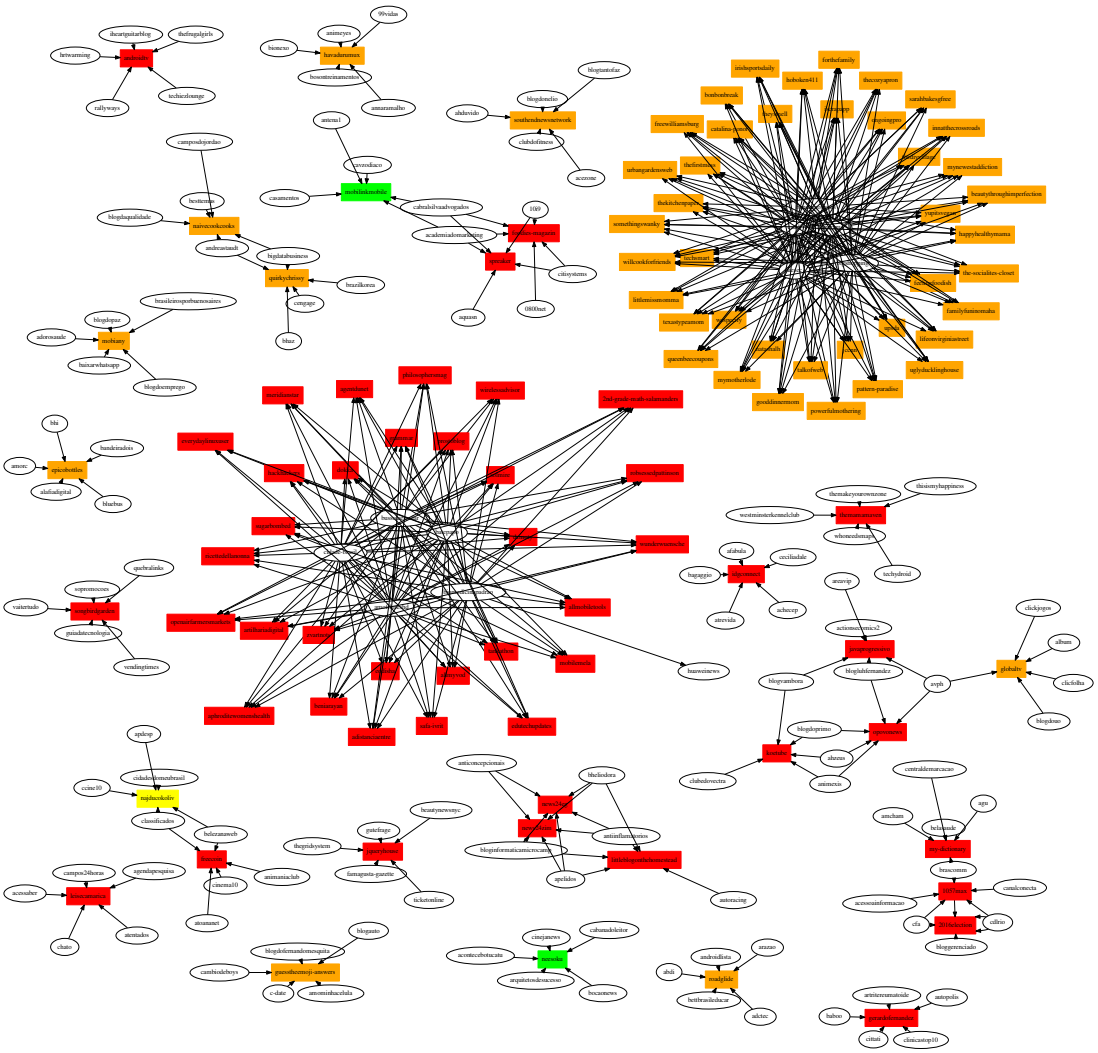

Figura 5.18: Grafo do Nível de Similaridade Base Alexa x 100 Consultas

.

Em outra perspectiva, a Figura5.19 destaca que 41 das 100 consultas relacionando-se com apenas 5 scripts da base de dados Alexa, dos quais destacam-se: http://

www.agorams.com.br, http://aiesec.org.br, http://bigshopping.com.br e http://www.caadf.org.br.

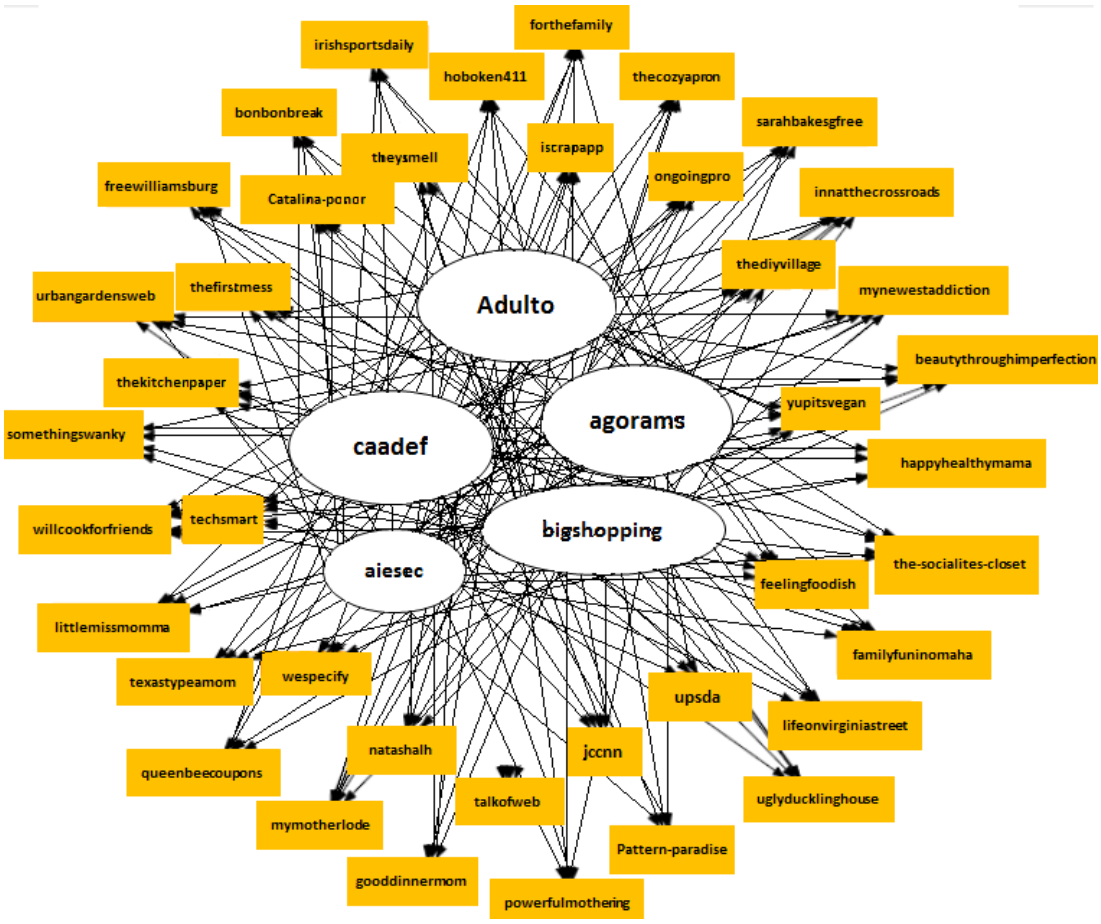

Figura 5.19: Parte do Grafo com o Resultado do Nível de Similaridade Base Alexa x 100 Consultas

.

É válido mencionar que para este experimento, obteve-se um resultado mediano em relação ao nível altíssimo e ao nível alto de similaridade em relação às 100 consultas, na imagem é prevalente as cores vermelho e laranja que representam um valor entre 90% a 100% e 80% a 89% de similaridade respectivamente, na comparação entre a base de dados Alexa e as 100 consultas.

Como forma de ilustrar os resultados descritos no cenário 3, a próxima seção 5.3.6, expõe cada um dos valores obtidos para as bases de dados em questão.

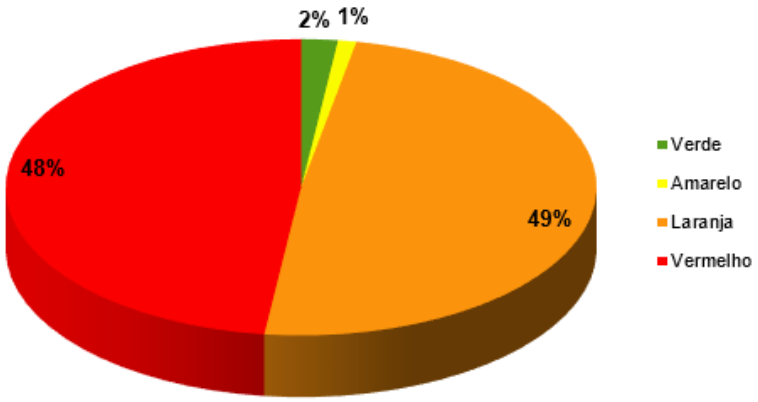

Figura 5.20: Gráfico do Nível de Similaridade Base Alexa x 100 Consultas

.

5.3.6 Discussão do Cenário 3

Em relação ao terceiro cenário, a Tabela 5.9 apresenta, de maneira sucinta, uma comparação entre os resultados obtidos nas 4 (quatro) base de dados, elucidando os níveis percentuais de similaridade obtidos em cada uma.

Tabela 5.9: Nível Percentual de Similaridade

Nível Percentual de Similaridade					
Base	**Qt. Sites**	**Baixo**	**Médio**	**Alto**	**Altíssimo**
Canvas	8.000	0%	1%	2%	**97%**
Phishtank	2.050	2%	6%	5%	**87%**
DMOZ	596	1%	8%	4%	**87%**
Alexa	1.478	2%	1%	**49%**	**48%**

É possível notar que a maioria das bases obtiveram os maiores níveis de scripts de sites com Altíssima similaridade. Entretanto a Canvas, apresentou o maior com 97%. A explicação para tal fato nesta base (Canvas) é até certo ponto simples, pois a mesma é composta realmente por sites relacionados (rotulados) com *fingerprinting*. Ou talvez, porque as consultas são provenientes desta própria base de dados. Já a segunda base Phishtank, dita como uma base de sites de *phishing*, também destaca um nível de similaridade altíssimo, com 87%, que de certa forma esperava-se por se tratar de uma base de sites maliciosos.

Porém, o resultado "inesperado" deu-se com as bases DMOZ e Alexa, visto que a primeira é composta por sites ditos benignos e a segunda é composta por sites hospedados no Brasil ligados a atividades como entretenimento e e-commerce,

por exemplo, obtendo-se para a primeira um nível Altíssimo com 87% e para a segunda as maiores partes da base dividiram-se entre basicamente ao meio com o nível Alto cerca de 49% e Altíssimo cerca de 48%, ou seja os valores somados incidiriam em 97% de similaridade, o mesmo valor percentual da base Canvas.

Como forma de validar todos os experimentos descritos em cada um dos cenários expostos anteriormente, a próxima seção 5.4, destaca três modos de validar os experimentos explicitados nestes cenários.

5.4 Validação

Para validar o método proposto, a seguir tem-se duas maneiras de validação: a primeira através do trabalho de Saraiva [30] e a segunda por meio da contagem das 41 características propostas nessa pesquisa.

5.4.1 Validação para Duas Características Canvas *Fingerprinting*

Para esta validação utilizou-se o trabalho proposto por Saraiva [30] o qual faz uma classificação de severidade de risco de ataques *fingerprinting* em três níveis: baixo, médio e alto. Para o nível alto, a autora destaca duas propriedades Canvas *fingerprinting* de alta periculosidade: $Canvas.toDataUrl()$ e $Canvas.getImageData()$. O método $Canvas.getImageData()$ retorna um objeto *ImageData* que copia os dados de pixel para o retângulo especificado em área Canvas. Na verdade, é preciso esclarecer que um objeto *ImageData* não é uma figura, e sim parte da área Canvas e manipula a informação de cada pixel dentro daquele retângulo. Já o método $Canvas.toDataURL()$ permite obter o conteúdo da tela do navegador. Os dados retornados formam uma string que representa uma URL codificada que contém os dados gráficos. Sendo estes de alto risco para obtenção de dados dos usuários Web.

Desta forma, esta pesquisa aplica somente estas duas características no método proposto, para validar a ideia e os resultados são os seguintes:

A Figura 5.21 mostra que para cada base de dados há incidências das duas características Canvas ($Canvas.toDataUrl()$ e $Canvas.getImageData()$), onde

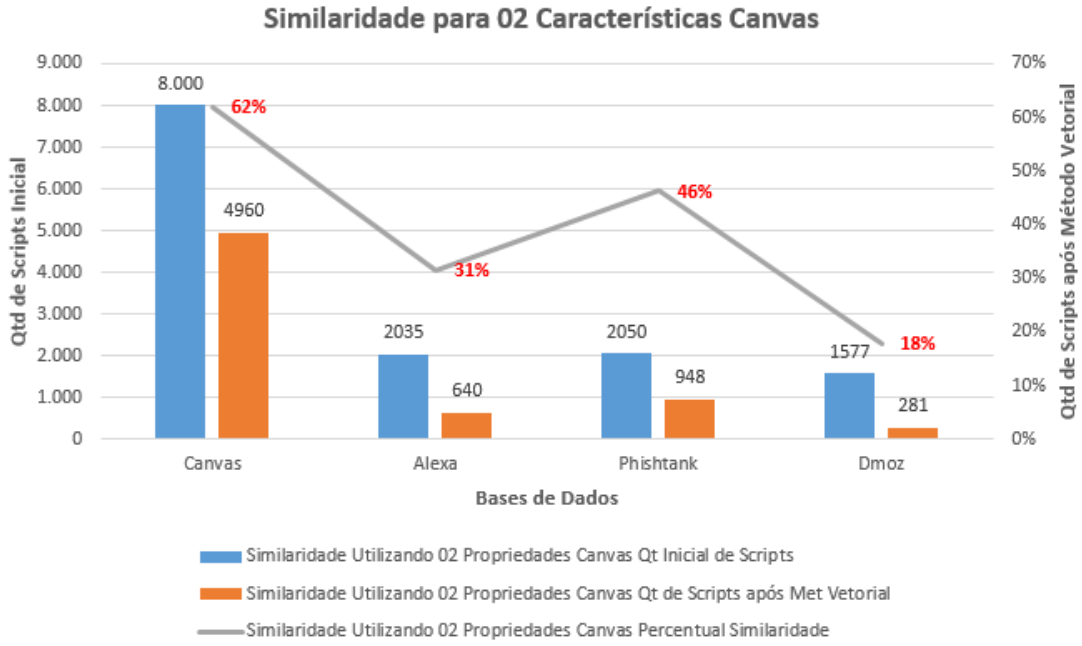

Figura 5.21: Gráfico do Nível de Similaridade para Duas Características Canvas

.

para a base Canvas: dos 8.000 scripts iniciais, ao aplicar o método proposto utilizando as duas características obteve-se um resultado onde 4.960 scripts continham as características de alta periculosidade. Para a base Alexa: 2.035 scripts iniciais, ao aplicar o método proposto utilizando as duas características obteve-se um resultado onde 640 scripts que possuíam as características propostas por Saraiva [30]. Na base Phishtank, dos 2.050 scripts iniciais, 948 destes scripts tinham as características. E para a base DMOZ, dos 1.577 scripts iniciais, cerca de 281 scripts foram detectados contendo estas duas características. Os resultados representam em pontos percentuais: 61%, 31%, 46% e 18% respectivamente. Em outras palavras, a ideia proposta nesta pesquisa foi validada, no que tange o uso das características Canvas *fingerprinting* nos scripts das páginas web, visto que as duas características elucidadas por Saraiva foram detectadas em todas as bases de dados, incluindo a benigna.

A outra maneira de validação do método por meio da contagem das características é explicitada na seção 5.4.2.

5.4.2 Validação Através da Contagem das Características Canvas *fingerprinting*

Uma outra maneira de validar o método proposto é apresentada por meio da contagem das características, para tanto, com base na Figura 5.10, elaborou-se a Tabela 5.10, a qual destaca os scripts da base Canvas que tem relação com 17 scripts das consultas. O resultado pode ser conferido a seguir:

Tabela 5.10: Contagem das Características Encontrados nos Scripts Canvas

Script	Contagem das Características pelo Método Proposto						
	fillText	textBaseline	clearRect	rotate	scale	getImageData	toDataURL
themakeyourownzone	54	6	12	0	0	30	18
techydroid	27	3	6	0	0	15	9
thisismyhappinees	27	3	6	0	0	15	9
Whoneedsmaps	27	3	6	0	0	15	9
Westminsterkennelclub	27	3	6	0	0	15	9
Script	**Contagem Manual das Características - Validação**						
	fillText	textBaseline	clearRect	rotate	scale	getImageData	toDataURL
themakeyourownzone	54	6	12	18	318	30	18
techydroid	27	3	6	0	21	15	9
thisismyhappinees	27	3	6	0	156	15	9
Whoneedsmaps	27	3	6	78	114	15	9
Westminsterkennelclub	27	3	6	78	114	15	9

A Tabela 5.10 apresenta variações nos valores de algumas das características entre a conferência pelo método proposto e a conferência manual, este fato ocorre, pois características como por exemplo: *rotate* e *scale*, que são presentes na API Canvas, mas também constam como funções de modificação de elementos na linguagem de folhas de estilo CSS.[4]

Assim como na tabela anterior, utilizando scripts da base de dados e das consultas, esta dissertação apresenta uma conferência semelhante a anterior, por meio de scripts maiores, com a finalidade de validar a ideia. Com isso, obteve-se os resultados expostos na Tabela 5.11:

Assim como nos resultados apresentados anteriormente, a Tabela 5.11 demonstra valores distintos para algumas das características, como por exemplo os dos scripts: 2016election que para a característica *lineTo* ressalta um valor de 64 no resultado do método e 68 para o resultado da conferência manual; outro

[4]Cascading Style Sheets (CSS) - Trata-se de uma linguagem de folhas de estilo que serve para definir a apresentação de documentos escritos em uma linguagem de marcação, como HTML ou XML.

Tabela 5.11: Conferência das Características

	Características Canvas Fingerprinting													
Scripts Canvas	fillText	fillStyle	textBaseline	strokeStyle	lineTo	lineWidth	textAlign	beginPath	restore	rotate	scale	strokeText	getImageData	toDataURL()
Contagem das Características pelo Método Proposto														
in-the-sky.org	166	162	112	150	218	126	126	192	34	12	90	26	18	0
2016election.com	36	42	12	46	64	46	26	48	10	16	4	0	10	6
theeducatorsroom.com	30	75	3	27	72	60	0	21	9	15	15	0	21	9
songbirdgarden.com	102	102	102	0	0	0	0	0	0	0	0	0	34	68
1057max.fm	72	108	52	100	148	104	80	108	68	44	16	4	24	4
Contagem Manual das Características - Validação														
in-the-sky.org	166	192	112	150	222	136	144	192	38	156	250	26	18	0
2016election.com	36	42	12	46	68	86	26	48	16	30	518	0	10	6
theeducatorsroom.com	30	75	3	27	72	60	6	21	57	24	504	0	21	9
songbirdgarden.com	102	102	102	0	0	0	0	0	0	0	3	0	34	68
1057max.fm	72	168	52	100	156	200	120	108	192	208	1968	4	24	4

script em que há divergência de valores é o 1057max que explicita resultados que diferem para as características *lineTo*, *lineWidth*, *textAlign*, *restore*, *rotate* e *scale*, os resultados esta em desacordo, pois os valores encontrados na conferência manual traz números de contagem para qualquer palavra que contenha parte do nome de uma característica, com isso, os resultados poderão ser maiores, visto que alguns dos métodos/propriedades Canvas, também podem ser encontrados em outras tecnologias como o CSS. Um outro fator importante a ser considerado é a questão de comentários, ou chamada de funções que não pertencem ao Canvas. Portanto, o método proposto só recupera informações do que realmente é importante, ou seja, a propriedade/método Canvas.

5.5 Discussão da Validação

Em relação aos aspectos da validação dos dados, esta dissertação ratifica que com as três maneiras de validar, sendo a primeira utilizando as características mencionas por Saraiva [30], a citar ($Canvas.toDataUrl()$ e $Canvas.getImageData()$), das quais a autora afirma serem de alta periculosidade, os resultados em destaque na seção 5.4.1, atesta que para as bases Canvas, Phishtank, DMOZ e Alexa, obteve-se os níveis de similaridade: 62%, 46%, 18% e 31%, respectivamente. Neste resultado, a base Canvas destaca-se com o nível altíssimo de

similaridade, visto que é a base composta por sites *fingerprinting* e comporta-se da mesma maneira que a base anterior, mesmo para a conferência com duas características. Nas demais bases, há uma redução de valores, pois procura-se apendas duas características, com isso há uma diminuição na amostra de scripts das bases de dados.

No que tange as duas últimas validações, ambas relacionadas à conferência das características, seja pelo método proposto ou de modo manual, as tabelas 5.10 e 5.11 apresentam quatorze (14) características com os valores maiores da contagem para cada uma delas. E apesar de haver distinção de valores nas contagens pelo método e pelo modo manual, a diferença se dá, pois para alguns scripts as características procuradas pelo primeiro só recupera a contagem para o Canvas, quanto que a segunda, além de trazer valores para Canvas, também recupera valores para palavras relacionadas ao rótulo da característica, ou ainda busca pelo mesmo rótulo, mas para outras tecnologias como o CSS, ou ainda, faz parte de comentários encontrados nos scripts. Fato que deixa explícito que esta dissertação só recupera o que está relacionado ao Canvas *fingerprinting*, como, por exemplo, no script `http://themakeyourownzone.com` da Tabela 5.10, na seção 5.4.2 que para a característica *scale*, na conferência pelo método apresenta um quantitativo de zero(0) e na conferência manual obtém trezentos e dezoito (318), pelo motivo descrito anteriormente.

Um outro exemplo que pode-se citar consta na Tabela 5.11 na seção 5.4.2, por meio do script `http://2016election.com` que na maioria das características apresenta um valor similar para ambas as conferências, porém nas características: *lineTo*, *lineWidth*, *restore*, *rotate* e *scale*, há uma diferença entre as contagens, obtendo-se os valores: pelo método (64, 46, 10, 16 e 4) e pela contagem manual (68, 86, 16, 30 e 518), respectivamente. Pois o método só preocupa-se na busca pelas características Canvas e a contagem manual traz não somente os valores para Canvas, como também para palavras semelhantes, comentários e ainda para características com a mesma nomenclatura referentes a outras tecnologias utilizadas na elaboração de scripts web.

Diante do exposto, as três validações mencionadas anteriormente deixam explícito que o método proposto nesta pesquisa, o qual tem a finalidade de ranquear pelo nível de similaridade os scripts das bases de dados, realizando as comparações

com as consultas Canvas resultou em cálculos de similaridade e ranqueamento dos cinco scripts mais semelhantes as consultas, alcançando com isso a ideia inicial da pesquisa.

Capítulo 6

Considerações Finais

Nos últimos anos, a técnica de *fingerprinting* têm sido recorrente nas páginas Web, visto que muitas empresas de publicidade e propaganda tem repassado as informações coletadas dos sites para seus parceiros comerciais. No que tange a questão da privacidade, infelizmente os usuários desconhecem este tipo de rastreamento de dados. Apesar de diversos estudos e publicações que têm sido propagadas como forma de alertar e dirimir a possibilidade da coleta de dados dos usuários da Internet, o fato ainda é bastante recorrente.

Assim como a Internet proporciona diversos tipos de atividades comuns e de grande valia para seus usuários, por trás de muitas páginas Web (*scripts*) há sempre a possibilidade de obtenção de informações que podem causar danos a privacidade daqueles que só pretendem navegar para realizar os afazeres comuns do cotidiano (como pagar um conta on-line, acessar seu e-mail e etc.). Este fato impõe que os profissionais de informática busquem soluções para impedir/dirimir os casos de furto de informações, engenharia social, técnica de *Website Fingerprinting* e outros que poderão ser prejudiciais àqueles que utilizam-na.

Esta pesquisa propôs um método para analisar *scripts* de Canvas *fingerprinting* em páginas Web e informar aos usuários o nível de similaridade entre a página e as consultas que poderão ser prejudiciais a sua privacidade. Deste modo, o conjunto de características (propriedades e métodos) Canvas *fingerprinting* foram avaliados e o nível de similaridade foi calculado.

Com isso, nesta pesquisa várias bases de dados foram averiguadas sendo esta,

uma maneira de provar, a existência do Canvas *fingerprinting* nos *scripts* das páginas web analisadas. Por meio de testes, em três cenários, os resultados demonstram que há um alto nível de similaridade entre as quatro bases de dados e as consultas Canvas.

6.1 Dificuldades encontradas

A primeira dificuldade a qual pode-se relatar, pauta-se na questão do download dos scripts das bases de dados, que desde meados de 2015 já havia sido realizado, porém notou-se que a instabilidade da rede de dados, fazia com que alguns scripts retornassem um valor menor do que seu tamanho original. Por exemplo, um script da base Alexa que fora feito download possuía um tamanho original de 1Mb, porém ao executar o download este retornava um tamanho de 600Kb, ou seja, 400Kb a menos de seu tamanho original. Fato que solicitou um retrabalho para cada uma das bases de dados.

Já em relação a ideia da detecção de scripts *fingerprinting* em páginas Web, inicialmente fora pensada a utilização dos algoritmos de aprendizagem de máquina para classificá-los como *scripts fingerprinting* e *scripts* não *fingerprinting*. Entretanto, notou-se que a maioria das páginas web, utiliza-se das propriedades e métodos de diversas tecnologias, como o $Window$, *screen* e outros, que tem a finalidade de realizar ajuste no conteúdo das páginas, como forma de melhorar a qualidade de sua apresentação no dispositivo dos usuários, fato que proporcionou uma classificação errônea destes scripts.

Uma outra tentativa de encaminhar esta pesquisa enveredou-se através da aplicação da técnica de episódios frequentes para alertar os usuários em relação aos scripts *fingerprinting*, porém notou-se durante os estudos que este tipo de técnica é mais utilizada para a detecção de tráfego anômalo. Apesar da ideia ser bastante interessante, teria que dispor de um tempo maior para percorrer esta direção.

Com base na ideia de episódios frequentes, a pesquisa enveredou para a descoberta de episódios frequentes em textos. Inicialmente seria o ideal, já que as propriedades e métodos de *fingerprinting* são scripts (texto), porém, após o primeiro experimento, viu-se que o algoritmo tinha uma complexidade de $2^{(mxn)}$,

ou seja, para 10 scripts analisados, o tempo de execução fora superior a dez (10) horas, imagine o tempo que precisaria dispor para executar a maior base de dados Canvas que possui 8.000 scripts, fato que inviabilizou a utilização desta técnica.

Assim como outras ideias surgiram e acabaram tornando-se inviáveis, notou-se que a busca por scripts *fingerprinting* ou não *fingerprinting* poderia ser mais importante se levasse aos usuários o quão similar é um script de um site visitado em relação às bases de páginas contendo Canvas.

6.2 Trabalhos Futuros

A comunidade de segurança no âmbito mundial, frequentemente têm realizado pesquisas sobre *fingerprinting* por meio de novos experimentos, proporcionando não só a sociedade, mas principalmente a comunidade científica inovações tecnológicas. Assim, esta pesquisa, elenca alguns trabalhos futuros como forma de retomada do assunto, conforme destaque a seguir:

1. Utilizar técnicas de entropia para verificar as características (propriedades e métodos) que possuam uma maior relevância para a detecção de *fingeprinting*;

2. Manter e categorizar uma base de scripts para disponibilizar a comunidade acadêmica nos seus estudos referentes ao *fingerprinting*.

3. Propor um mecanismo para verificar características de *fingerprinting* ofuscadas em sites Web;

Referências Bibliográficas

[1] The Wall Street Journal, "Latest in web tracking: Stealthy supercookies," 2011.

[2] D. Kristol and L. Montulli, "HTTP State Management Mechanism." RFC 2109 (Proposed Standard), February 1997. http://www.ietf.org/rfc/rfc2109.txt.

[3] E. Flood and J. Karlsson, "Browser Fingerprinting," no. May, 2012.

[4] P. Ximenes, M. Correia, P. Mello, F. Carvalho, M. Franklin, and R. Andrade, "TARP Fingerprinting: Um Mecanismo de Browser Fingerprinting Baseado em HTML5 Resistente a Contramedidas," 2016.

[5] J. R. Mayer, "Internet Anonymity in the Age of Web 2.0," *A Senior Thesis presented to the Faculty of the Woodrow Wilson School of Public and International A airs in partial ful llment of the requirements for the degree of Bachelor of Arts.*, p. 103, 2009.

[6] P. Eckersley, "How unique is your web browser?," in *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS'10, (Berlin, Heidelberg), pp. 1–18, Springer-Verlag, 2010.

[7] A. Cooper, H. Tschofenig, B. Aboba, J. Peterson, J. Morris, M. Hansen, and R. Smith, "Privacy Considerations for Internet Protocols." RFC 6973 (Informational), July 2013. http://www.ietf.org/rfc/rfc6973.txt.

[8] W3C, "Fingerprinting guidance for web specification authors," 2014. http://w3c.github.io/fingerprinting-guidance/.

[9] A. Barth, "HTTP State Management Mechanism." RFC 6265 (Proposed Standard), Apr. 2011. http://www.ietf.org/rfc/rfc6265.txt.

[10] S. Kamkar, "Evercookie." http://samy.pl/evercookie/, 2010.

[11] K. Mowery and H. Shacham, "Pixel Perfect : Fingerprinting Canvas in HTML5," *Web 2.0 Security & Privacy 20 (W2SP)*, pp. 1–12, 2012.

[12] E. L. Saraiva, Adriana Rodrigues; Elleres, Pablo Augusto da Paz; Carneiro, Guilherme de Brito; Feitosa, "Device Fingerprinting: Conceitos e Técnicas, Exemplos e Contramedidas," Minicursos do XIV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais ? SBSeg 2014, 2014 ed., 2014.

[13] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, pp. 674–689, 2014.

[14] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, no. 1, pp. 1388–1401, 2016.

[15] M. Duarte, "Entropia, redundância... quanta informação, que confusão!." http://incomuniq.blogspot.com.br/2011/10/entropia-redundancia-quanta-informacao.html, 2011.

[16] A. F. Khademi, "Browser Fingerprinting : Analysis , Detection , and Prevention at Runtime," no. October, 2014.

[17] J. Kirk, "Canvas fingerprinting online tracking is sneaky but easy to halt." http://www.pcworld.com/article/2458280/canvas-fingerprinting-tracking-is-sneaky-but-easy-to-halt.html, 2014. [Online, Acessado 21/09/2014].

[18] R. &. R.-N. Baeza-Yates, *Recuperação de Informação - Conceitos e Tecnologia das Máquinas de Busca*. 2013.

[19] A. M. de. Oliveira, "Um método de detecção de plágio em códigos-fonte para disciplinas iniciais de programação," p. 84, 2016.

[20] R. S. Chaves, "Análise em agrupamentos de documentos eletrônicos," p. 78, 2014.

[21] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle"in a haystack: Detecting knee points in system behavior," *Proceedings - International Conference on Distributed Computing Systems*, pp. 166–171, 2011.

[22] D. Jang, R. Jhala, S. Lerner, and H. Shacham, "An empirical study of privacy-violating information flows in JavaScript web applications," *Proceedings of the 17th ACM conference on Computer and communications security - CCS '10*, p. 270, 2010.

[23] S. Boda, Károly; Földes, Ádám Máté; Gulyás, Gábor György; Imre, "User Tracking on the Web via Cross-Browser," *Proceeding NordSec'11 Proceedings of the 16th Nordic conference on Information Security Technology for Applications Pages 31-46*, pp. 1–17, 2012.

[24] T.-f. Yen, Y. Xie, F. Yu, and R. P. Yu, "Host Fingerprinting and Tracking on the Web : Privacy and Security Implications," 2012.

[25] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel, "FPDetective : Dusting the Web for Fingerprinters Categories and Subject Descriptors," 2013.

[26] L. Olejnik, C. Castelluccia, and A. Janc, "On the uniqueness of Web browsing history patterns," *Annals of Telecommunications - Annales Des Télécommunications*, vol. 69, pp. 63–74, Sept. 2013.

[27] P. Laperdrix, W. Rudametkin, and B. Baudry, "Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints," *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pp. 878–894, 2016.

[28] E. Bursztein, A. Malyshev, T. Pietraszek, and K. Thomas, "Picasso: Lightweight Device Class Fingerprinting for Web Clients," *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices*, pp. 93–102, 2016.

[29] G. Nakibly, G. Shelef, and S. Yudilevich, "Hardware Fingerprinting Using HTML5," *Computing Research Repository (CoRR)*, vol. abs/1503.0, 2015.

[30] A. R. Saraiva, *Determinando o Risco de Fingerprinting em Páginas Web*. Dissertação de mestrado, Universidade Federal do Amazonas, 2016.

[31] GlobalAD, "O que é Crawler?," 2013.

[32] T. A. Sudkamp, *Introduction to the Second Edition*. 2 ed. ed., 2005.

[33] N. Nikiforakis, a. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting," *2013 IEEE Symposium on Security and Privacy*, pp. 541–555, May 2013.

[34] A. E. Nunan, *Detecção de Cross-Site Scripting em Páginas Web*. PhD thesis, 2012.

[35] Amazon, "Alexa." http://www.alexa.com/, 2017.

[36] ISO27000, "Classificação de Risco - ISO27000." http://iso27000.com.br, 2017.

Apêndice A

Resultados Canvas

a) Similaridade entre Scripts - Base Canvas x 100 Consultas

Tabela A.1: Similaridade entre Scripts (Base Canvas x 100 Consultas)

Similaridade entre Scripts (Base Canvas x 100 Consultas)			
Agrupamento de Scripts Base Canvas	**Scripts Base Canvas**	**Qtd Scripts no Grupo**	**Qtd Vezes *Ranking***
Grupo 1	**guidingtech, ftlauderdalewebcam, preppyrunner, embassypages, nzhuntingandshooting**	**5**	**29**
Grupo 2	**sandraandwoo, filmhafizasi, dcclothesline, cleverlyinspired**	**4**	**24**
Grupo 3	**eastcoastcreativeblog**	**1**	**23**
Grupo 4	thisismyhappiness, westminsterkennelclub, whoneedsmaps, techydroid, themakeyourownzone	5	17
Grupo 5	animexis, freebiesjedi, urdunovels, pistolsfiringblog, take-me-to-auction, shitlicious, outbreaknewstoday, criadoresdeconteudo	8	2
Grupo 6	bradfordcityfc, thedieselstop, thehatlogic, graciefrancesca, destinydb, vendingtimes, veteranstoday,willtirando, madetobeamomma, equinenow, fotograpiya, splashmagazine, freeonlinephotoeditor,noticiasparaiba, clubtouareg, soccer-douga, sopromocoes, itallstartedwithpaint, pixelshut, namebright, sa-mp, beautynewsnyc, mundodasmarcas, bitcoinfaucetexchange, laughspin, donandroid, ireport,booshsports, inati0n, juwelo, savingmoneylivingsmart, rallyways,lexibites, platinumblackcomic,hombremoderno, tamilmagazines, iheartguitarblog, techiezlounge, crisguerra,ticketcorner, videoblocks, 11magnolialane, sheboyganpress, tohapi, concertpass, nakluky, techranker, devzum, staffordshirenewsletter, patchworkposse, theurbantwist, c-date, trulygeeky, novinky, securitymagazin,thegridsystem, fdlreporter, guiadatecnologia, ligadoemserie, myschool, hearteating,youknowyoulovefashion,typingtestnow, piecesvoitures, fizzypeaches, nsmb, a1-forum, reidl,empreendedordigital, akdirahost, relyonhorror,noitesinistra, sigalert, fastsocialfollower, thesavvybump,folhetimonline, mixanitouxronou, blogvambora, blogdohammer,kellyinthecity, manaserials,pointemagazine, matomame, point-island, trendingtoplists, jmunicipios, prehled-prace, midiaesportiva,viladoesporte, rcmania, tap-repeatedly, gutefrage, freebtcfaucet, ch10, vaitertudo, azgrabaplate,caminhandojunto, famagusta-gazette, thewhitebuffalostylingco, thefrugalgirls, mobilebeat, warriorsworld, quebralinks, thesummeryumbrella, kawasakininja300, invasaonerd, sanjorgeschool, clarkscondensed,kitchenstewardship,lamianuovaclio, ticketonline, sammichespsychmeds, mycoinads, burgmanusa,theglamoroushousewife, domosfera, somethingtostream, hrtwarming,eventim, thenester, lotustalk, osxarena, honnest, boxfaucet, lucrena, milicoponderao,rekbitcoin, grantspassweather, dancemagazine, fleurdeforce, zipix, thewisebaby, stream-a-ams1xx2sfcdnvideo5269,beanpanda, im-internetmarketing	135	1

b) Consultas mais Relevantes - Base Canvas x 100 Consultas

Tabela A.2: Consultas mais Relevantes (Base Canvas x 100 Consultas)

Consultas mais Relevantes (Base Canvas x 100 Consultas)			
Agrupamento de Consultas	**Consultas**	**Qt Scripts da Base Canvas Relevantes**	**Percentual de Relevância Consulta x Base Canvas**
Consulta 1	**southendnewsnetwork**	**4873**	**61%**
Consulta 2	**epicobottles**	**3691**	**46%**
Consulta 3	androidtv	2809	35%
Consulta 4	mobiany	2793	35%
Consulta 5	themamamaven, hoboken411, thediyvillage, powerfulmothering, the-socialites-closet, wespecify, ongoingpro, somethingswanky, queenbeecoupons, irishsportsdaily, talkofweb, jccnn, sarahbakesgfree, mymotherlode, thecozyapron, iscrapapp, forthefamily, upsda, texastypeamom, techsmart, thefirstmess, happyhealthymama, pattern-paradise, bonbonbreak, willcookforfriends, beautythroughimperfection, mynewestaddiction, littlemissmomma, lifeonvirginiastreet, freewilliamsburg, natashalh, theysmell, familyfuninomaha, thekitchenpaper, feelingfoodish, urbangardensweb, catalina-ponor, yupitsvegan, gooddinnermom, uglyducklinghouse	2772	35%
Consulta 6	innatthecrossroads	2770	35%
Consulta 7	quirkychrissy	2746	34%
Consulta 8	naivecookcooks	2660	33%
Consulta 9	roadglide	2361	30%
Consulta 10	havadurumux	1843	23%
Consulta 11	guesstheemoji-answers	1830	23%
Consulta 12	littleblogonthehomestead	1702	21%
Consulta 13	mobilinkmobile	1648	21%
Consulta 14	news24eg	1639	20%
Consulta 15	news24zim	1570	20%
Consulta 16	neesoku	1372	17%
Consulta 17	meridianstar, openairfarmersmarkets, dokka, wunderwuensche, artilhariadigital, eodisha, everydaylinuxuser, solmire, philosophersmag, demoty, mobilemela, safa-ivrit, grammar, 2nd-grade-math-salamanders, aphroditewomenshealth, ricettedellanonna, beniarayan, robsessedpattinson, edutechupdates, allmyvod, adistanciaentre, zvartnots, allmobiletools, wirelessadvisor, hackhackers, prostoblog, sugarbombed, tankathon, agentdunet	1223	15%
Consulta 18	idgconnect	1158	14%
Consulta 19	javaprogressivo	909	11%
Consulta 20	songbirdgarden	872	11%
Consulta 21	freecoin	856	11%
Consulta 22	gerardofernandez	800	10%
Consulta 23	opovonews	795	10%
Consulta 24	koetube	763	10%
Consulta 25	my-dictionary	679	8%
Consulta 26	leisecamarica	552	7%
Consulta 27	globaltv	513	6%
Consulta 28	1057max	349	4%
Consulta 29	foodies-magazin	347	4%
Consulta 30	jqueryhouse, 2016election, spreaker	324	4%
Consulta 31	najducokoliv	107	1%

Apêndice B

Resultados Phishtank

a) Similaridade entre Scripts - Base Phishtank x 100 Consultas

Tabela B.1: Similaridade entre Scripts (Base Phishtank x 100 Consultas)

Similaridade entre Scripts (Base Phishtank x 100 Consultas)			
Agrupamento de Scripts Base Phishtank	**Script Base Phishtank**	**Qtd Scripts no Grupo**	**Qtd Vezes *Ranking***
Grupo 1	**caraudioacapulco**	**1**	**41**
Grupo 2	**sigarabirakmak**	**1**	**40**
Grupo 3	**info-setting2016, ricardoeletro2, infobel, maisponto, lagosstatenews**	**5**	**29**
Grupo 4	**replacementroofingtx, rcacas, davinciresidence**	**3**	**24**
Grupo 5	jimmyseas	1	18
Grupo 6	pasarpedia, pracadarepublicaembeja	2	17
Grupo 7	greenhub	1	7
Grupo 8	bnpparibas-abstract-expressionism, as	2	6
Grupo 9	mali-dugi	1	5
Grupo 10	promolinks	1	4
Grupo 11	kitchensrus	1	3
Grupo 12	ba-tango, en, lioninthesun, dotartprinting, secure-banklogin, pejdah-pharmacia, courchevel,clermontlounge, syengage, theworkouts, worldsalsachampionships, machizo, ibooking, becgiemob, ibusinessolution, ar, progorod, pallascaaldia	18	2
Grupo 13	pousadacavalinno, buildonlinewealth, beast-the-animal, wisefellas, footspecialistbrampton, twinyoubethinking, yonearts, ronarhost, esivah, capalaroche, nolagroup, mashavrelocation, fanatiksport, 911beautystudios, carverslaw, hackspirit, ipkill, itpbacau, familiatuccini, novatekit, bancofotografias, mybeltz, altavistawines, mbu, specialchild, innmoema, nubeviajera, correodegmail, t-alquds, eddaturkey, santechno, tra, banksignin, plasticmonkey, sportyfit, capitalcu, jehansen, autoscriba, lcmaquinasseladoras,elogix, stevewarshak, evenpro, choicesunlimited, jamiexxperth, serrauquer, clean-clean-peru, promowear, create-new-account, hyperbarichealingcenter, loginsign, hoop4eva, dianagarces, forexmoneyback, batallonchacras, danielabrantes, unitygamesbox, linkedin, couponshe, moneyexprt, nadiabernocchi, aliexpress, giftcardblogger, lilicoimoveis, dhaainkanbaa, globalcleaning, es, wibs, goo, flatbellyfitness, keeneteamvegas, healthinsurance, sire-china, wistub-brenner, eurochistka, kazichschool, hism, spotcampus, eclecticgrape, farlainlake, beffy, biovene, notafog2016, myhotmailsignin	83	1

b) Consultas mais Relevantes - Base Phishtank x 100 Consultas

Tabela B.2: Consultas mais Relevantes (Base Phishtank x 100 Consultas)

Consultas mais Relevantes (Base Phishtank x 100 Consultas)			
Agrupamento de Consultas	**Consultas**	**Qt Scripts da Base Phishtank Relevantes**	**Percentual Relevância Consulta x Base Phishtank**
Consulta 1	**southendnewsnetwork**	**975**	**48%**
Consulta 2	**epicobottles**	**843**	**41%**
Consulta 3	guesstheemoji-answers	800	39%
Consulta 4	mobilinkmobile	625	30%
Consulta 5	gerardofernandez	541	26%
Consulta 6	idgconnect	526	26%
Consulta 7	mobiany	513	25%
Consulta 8	roadglide	511	25%
Consulta 9	quirkychrissy	504	25%
Consulta 10	naivecookcooks	492	24%
Consulta 11	innatthecrossroads	491	24%
Consulta 12	themamamaven, hoboken411, thediyvillage, powerfulmothering, the-socialites-closet, wespecify, ongoingpro, somethingswanky, queenbeecoupons, irishsportsdaily, talkofweb, jccnn, sarahbakesgfree, mymotherlode, thecozyapron, iscrapapp, forthefamily, texastypeamom, upsda, techsmart, thefirstmess, happyhealthymama, pattern-paradise, bonbonbreak, willcookforfriends, beautythroughimperfection, mynewestaddiction, littlemissmomma, lifeonvirginiastreet, freewilliamsburg, natashalh, theysmell, familyfuninomaha, thekitchenpaper, feelingfoodish, urbangardensweb, catalina-ponor, yupitsvegan, gooddinnermom, uglyducklinghouse	489	24%
Consulta 13	androidtv	479	23%
Consulta 14	littleblogonthehomestead	437	21%
Consulta 15	news24eg	414	20%
Consulta 16	news24zim	408	20%
Consulta 17	1057max	255	12%
Consulta 18	2016election	243	12%
Consulta 19	havadurumux	218	11%
Consulta 20	globaltv	197	10%
Consulta 21	my-dictionary	187	9%
Consulta 22	foodies-magazin	178	9%
Consulta 23	najducokoliv	170	8%
Consulta 24	spreaker	154	8%
Consulta 25	freecoin	153	7%
Consulta 26	meridianstar, openairfarmersmarkets, dokka, wunderwuensche, artilhariadigital, eodisha, everydaylinuxuser,solmire, philosophersmag, demoty, mobilemela, safa-ivrit, beniarayan, 2nd-grade-math-salamanders, allmyvod,aphroditewomenshealth, ricettedellanonna, robsessedpattinson, edutechupdates, grammar, zvartnots, allmobiletools, wirelessadvisor, hackhackers, prostoblog, sugarbombed, tankathon, agentdunet, adistanciaentre	143	7%
Consulta 27	jqueryhouse	125	6%
Consulta 28	songbirdgarden	94	5%
Consulta 29	javaprogressivo	89	4%
Consulta 30	neesoku, koetube	66	3%
Consulta 31	opovonews	65	3%
Consulta 32	leisecamarica	16	1%

Apêndice C

Resultados DMOZ

a) Similaridade entre Scripts - Base Dmoz x 100 Consultas

Tabela C.1: Similaridade entre Scripts (Base Dmoz x 100 Consultas)

Similaridade entre Scripts (Base Dmoz x 100 Consultas)			
Agrupamento de Scripts Base Dmoz	**Script Base Dmoz**	**Qt Scripts no Grupo**	**Qtd Vezes Ranqueada**
Grupo 1	**mypet-memorial**	**1**	**44**
Grupo 2	**sonicyoga, modern-rocket, mdsafrica**	**3**	**43**
Grupo 3	**panta-rhei**	**1**	**42**
Grupo 4	**bobthealien, cottonclouds, shesmoke, plum**	**4**	**29**
Grupo 5	madmeatgenius	1	16
Grupo 6	highpowergraphics	1	13
Grupo 7	23d	1	7
Grupo 8	itworld, networkworld	2	6
Grupo 9	renewingthecountryside	1	4
Grupo 10	aidsinfo, jnproductions, flairstrips, 26and2, ker, goalballuk, petmd, earthdance, prenatalyogacenter, himalayaninstitute, accentonline, onlinelibrary, pancero	13	3
Grupo 11	paintballtraining, apartment-ideas, eplbas, simonsmd, ccvideo, smoking-meat, countrysidenetwork, cloudbyuchit, cowgirlscountry, cbrdigital, kingsford, abcnews, farewellfurryfriend, plausiblefutures, atmananda, ontariogenomics, barbequemaster	17	2
Grupo 12	mas, grillgirl, kodak, safekids, horizonvp, manningdigital, stillpointyogastudios, softcat, dellarte, cyndilee, robotshop, bbqsmokersite, cambo, homebbq, fezana, accessgenealogy, stopthefleas, fibergarden, petlossjourney, bhavayoga, doublezoot, howtobbqright, productioncraft, ellenbogen, gaetanomansi, andreschuster, robinsonarchive, kidoz, advmediaservices, houstonyoga, asr-gooyesh, leaguefreak, cosmoquest, herrschners, rjmuna, accompanyvideo, grillingandsmoking, astronomicaloptical, beyondtrust, richiegillphotography, staylor-made, kamadojim, biggreeneggic, lifeinmotion	44	1

b) Consultas mais Relevantes - Base Dmoz x 100 Consultas

Tabela C.2: Consultas mais Relevantes (Base Dmoz x 100 Consultas)

Consultas mais Relevantes (Base Dmoz x 100 Consultas)		
Agrupamento de Consultas	**Consultas**	**Qt Scripts Base Dmoz Relevantes**
Consulta 1	**southendnewsnetwork**	**339**
Consulta 2	**quirkychrissy**	**258**
Consulta 3	idgconnect	195
Consulta 4	guesstheemoji-answers	193
Consulta 5	news24eg	166
Consulta 6	littleblogonthehomestead	164
Consulta 7	news24zim	163
Consulta 8	havadurumux, leisecamarica	157
Consulta 9	epicobottles	156
Consulta 10	themamamaven, hoboken411, thediyvillage, powerfulmothering, the-socialites-closet, wespecify, ongoingpro, innatthecrossroads, somethingswanky, queenbeecoupons, irishsportsdaily, talkofweb, jccnn, sarahbakesgfree, mymotherlode, thecozyapron, iscrapapp, forthefamily, texastypeamom, upsda, techsmart, thefirstmess, happyhealthymama, pattern-paradise, bonbonbreak, willcookforfriends, beautythroughimperfection, mynewestaddiction, littlemissmomma, freewilliamsburg, natashalh, theysmell, familyfuninomaha, thekitchenpaper, feelingfoodish, urbangardensweb, catalina-ponor, yupitsvegan, gooddinnermom, uglyducklinghouse	152
Consulta 11	androidtv	151
Consulta 12	mobiany	146
Consulta 13	naivecookcooks	137
Consulta 14	mobilinkmobile	134
Consulta 15	roadglide	114
Consulta 16	gerardofernandez	97
Consulta 17	1057max	93
Consulta 18	2016election	89
Consulta 19	foodies-magazin	82
Consulta 20	my-dictionary, najducokoliv	72
Consulta 21	neesoku, spreaker	51
Consulta 22	meridianstar, openairfarmersmarkets, dokka, wunderwuensche, artilhariadigital, eodisha, everydaylinuxuser, solmire, philosophersmag, demoty, mobilemela, safa-ivrit, 2nd-grade-math-salamanders, aphroditewomenshealth, ricettedellanonna, beniarayan, robsessedpattinson, edutechupdates, allmyvod, adistanciaentre, grammar, zvartnots, allmobiletools, wirelessadvisor, hackhackers, prostoblog, sugarbombed, tankathon, agentdunet	49
Consulta 23	jqueryhouse	36
Consulta 24	javaprogressivo	35
Consulta 25	globaltv	27
Consulta 26	freecoin	26
Consulta 27	songbirdgarden	24
Consulta 28	koetube	16
Consulta 29	opovonews	15

Apêndice D

Resultados Alexa

a) Similaridade entre Scripts - Base Alexa x 100 Consultas

Tabela D.1: Similaridade entre Scripts (Base Alexa x 100 Consultas)

Similaridade entre Scripts (Base Alexa x 100 Consultas)			
Agrupamento de Scripts Base Alexa	**Script Base Alexa**	**Qt Scripts no Grupo**	**Qtd Vezes Ranqueada**
Grupo 1	**bundasgostosas, caadf, agorams, aiesec, bigshopping**	**5**	**41**
Grupo 2	**biomedicinapadrao, amofilmeshd, batepapo, cidade-brasil, bussolaescolar**	**5**	**29**
Grupo 3	avph, academiadomarketing, cabralsilvaadvogados	3	4
Grupo 4	blogluhfernandez, 10i9, bheliodora, antiinflamatorios, citisystems, apelidos, animexis, bloginformaticamicrocamp, brascomm	9	3
Grupo 5	blogdouo, cfa, cdlrio, ahzeus, blogdoprimo, aquasn, andreastaudt, clicfolha, bigdatabusiness, belezanaweb, anticoncepcionais, acessoainformacao, classificados, blogvambora	14	2
Grupo 6	acezone, abdi, andrezadicaeindicadisney, arazao, brazilkorea, 0800net, clinicastop10, agoramt, blogdaqualidade, baboo, bionexo, bluebus, cittati, bloggerenciado, afabula, clubdofitness, amominhacelula, centraldemarcacao, acontecebotucatu, amorc, cidadesdomeubrasil, artritereumatoide, animaniaclub, blogtantofaz, amcham, baixarwhatsapp, alemdaruaatelier, canalconecta, casamentos, clickcamboriu, 1001cupomdedescontos, bosontreinamentos, agendapesquisa, atrevida, cinema10, actionsecomics2, annaramalho, belasaude, alafiadigital, cabanadoleitor, animeyes, camposdojordao, autopolis, album, cengage, arquitetosdesucesso, cambiodeboys, 99vidas, bhi, androidlista, acessaber, clubedovectra, c-date, areavip, ahduvido, bandeiradois, campos24horas, agu, ceciliadale, brasileirosporbuenosaires, apdesp, blogdoemprego, blogdopaz, achecep, chato, atoananet, antena1, bocaonews, blogdofernandomesquita, bettbrasileducar, blogdonelio, adctec, bagaggio, bhaz,,blogauto, clickjogos, cinejanews, ccine10, autoracing, adorosaude, atentados, cavzodiaco, besttemas	83	1

b) Consultas mais Relevantes - Base Alexa x 100 Consultas

Tabela D.2: Consultas mais Relevantes (Base Alexa x 100 Consultas)

Consultas mais Relevantes (Base Alexa x 100 Consultas)			
Agrupamento de Consultas	**Consultas**	**Qt Scripts da Base Alexa Relevantes**	**Percentual de Relevância Consulta x Base Alexa**
Consulta 1	**quirkychrissy**	**620**	**42%**
Consulta 2	**naivecookcooks**	**595**	**40%**
Consulta 3	southendnewsnetwork	541	37%
Consulta 4	idgconnect	491	33%
Consulta 5	guesstheemoji-answers	424	29%
Consulta 6	littleblogonthehomestead	407	28%
Consulta 7	news24eg, news24zim	334	23%
Consulta 8	androidtv	294	20%
Consulta 9	mobiany	287	19%
Consulta 10	epicobottles	279	19%
Consulta 11	1057max	271	18%
Consulta 12	havadurumux	254	17%
Consulta 13	themamamaven, hoboken411, thediyvillage, powerfulmothering, the-socialites-closet, wespecify, ongoingpro, innatthecrossroads, somethingswanky, queenbeecoupons, irishsportsdaily, talkofweb, jccnn, sarahbakesgfree, mymotherlode, thecozyapron, iscrapapp, forthefamily, texastypeamom, upsda, techsmart, thefirstmess, happyhealthymama, pattern-paradise, bonbonbreak, willcookforfriends, beautythroughimperfection, mynewestaddiction, littlemissmomma, lifeonvirginiastreet, freewilliamsburg, natashalh, theysmell, familyfuninomaha, thekitchenpaper, feelingfoodish, urbangardensweb, catalina-ponor, yupitsvegan, gooddinnermom, uglyducklinghouse	251	17%
Consulta 14	my-dictionary	178	12%
Consulta 15	gerardofernandez	175	12%
Consulta 16	2016election	171	12%
Consulta 17	freecoin	162	11%
Consulta 18	najducokoliv	149	10%
Consulta 19	javaprogressivo	144	10%
Consulta 20	songbirdgarden, meridianstar, openairfarmersmarkets, dokka, wunderwuensche, artilhariadigital, eodisha, everydaylinuxuser, solmire, philosophersmag, demoty, mobilemela, safa-ivrit, 2nd-grade-math-salamanders, aphroditewomenshealth, ricettedellanonna, beniarayan, robsessedpattinson, edutechupdates, allmyvod, adistanciaentre, grammar, zvartnots, allmobiletools, wirelessadvisor, tankathon, hackhackers, prostoblog, sugarbombed, agentdunet	139	9%
Consulta 21	mobilinkmobile	120	8%
Consulta 22	foodies-magazin	114	8%
Consulta 23	globaltv	108	7%
Consulta 24	koetube	105	7%
Consulta 25	opovonews	104	7%
Consulta 26	spreaker	100	7%
Consulta 27	neesoku	89	6%
Consulta 28	roadglide	83	6%
Consulta 29	jqueryhouse	71	5%
Consulta 30	leisecamarica	65	4%

Apêndice E

Similaridade Top 5 Canvas

Tabela E.1: Tabela Similaridade Base Canvas x 100 Consultas - Parte1

Tabela Nível de Similaridade Base Canvas x 100 Consultas (Parte 1)					
Consulta	**Script Base Canvas**	**Similaridade**	**Consulta**	**Script Base Canvas**	**Similaridade**
themamamaven	techydroid	1.0	**epicobottles**	thesummeryumbrella	1.0
	whoneedsmaps	1.0		trendingtoplists	1.0
	westminsterkennelclub	1.0		thenester	1.0
	thisismyhappiness	1.0		thesavvybump	1.0
	themakeyourownzone	1.0		mobilebeat	1.0
androidtv	techiezlounge	0.999	**powerfulmothering**	filmhafizasi	1.0
	rallyways	0.999		sandraandwoo	1.0
	hrtwarming	0.999		cleverlyinspired	1.0
	thefrugalgirls	0.999		dcclothesline	1.0
	iheartguitarblog	0.999		eastcoastcreativeblog	1.0
jqueryhouse	ticketonline	1.0	**news24eg**	outbreaknewstoday	1.0
	gazette	0.997		take-me-to-auction	0.999
	thegridsystem	0.997		urdunovels	0.997
	beautynewsnyc	0.997		shitlicious	0.997
	gutefrage	0.997		freebiesjedi	0.997
songbirdgarden	vendingtimes	1.0	**dokka**	embassypages	1.0
	guiadatecnologia	0.994		preppyrunner	1.0
	vaitertudo	0.994		guidingtech	1.0
	quebralinks	0.994		nzhuntingandshooting	1.0
	sopromocoes	0.994		ftlauderdalewebcam	1.0
hoboken411	filmhafizasi	1.0	**wunderwuensche**	embassypages	1.0
	sandraandwoo	1.0		preppyrunner	1.0
	cleverlyinspired	1.0		guidingtech	1.0
	dcclothesline	1.0		nzhuntingandshooting	1.0
	eastcoastcreativeblog	1.0		ftlauderdalewebcam	1.0
meridianstar	embassypages	1.0	**artilhariadigital**	embassypages	1.0
	preppyrunner	1.0		preppyrunner	1.0
	guidingtech	1.0		guidingtech	1.0
	nzhuntingandshooting	1.0		nzhuntingandshooting	1.0
	ftlauderdalewebcam	1.0		ftlauderdalewebcam	1.0
havadurumux	fleurdeforce	1.0	**the-socialites-closet**	techydroid	1.0
	thehatlogic	0.998		whoneedsmaps	1.0
	fizzypeaches	0.996		westminsterkennelclub	1.0
	graciefrancesca	0.961		thisismyhappiness	1.0
	lamianuovaclio	0.958		themakeyourownzone	1.0

Tabela E.2: Tabela Similaridade Base Canvas x 100 Consultas - Parte2

Tabela Nível de Similaridade Base Canvas x 100 Consultas (Parte 2)					
Consulta	**Script Base Canvas**	**Similaridade**	**Consulta**	**Script Base Canvas**	**Similaridade**
openairfarmersmarkets	embassypages	1.0	**eodisha**	embassypages	1.0
	preppyrunner	1.0		preppyrunner	1.0
	guidingtech	1.0		guidingtech	1.0
	nzhuntingandshooting	1.0		nzhuntingandshooting	1.0
	ftlauderdalewebcam	1.0		ftlauderdalewebcam	1.0
thediyvillage	techydroid	1.0	**roadglide**	kawasakininja300	1.0
	whoneedsmaps	1.0		lotustalk	1.0
	westminsterkennelclub	1.0		burgmanusa	1.0
	thisismyhappiness	1.0		clubtouareg	1.0
	themakeyourownzone	1.0		thedieselstop	1.0
ongoingpro	filmhafizasi	1.0	**wespecify**	techydroid	1.0
	sandraandwoo	1.0		whoneedsmaps	1.0
	cleverlyinspired	1.0		westminsterkennelclub	1.0
	dcclothesline	1.0		thisismyhappiness	1.0
	eastcoastcreativeblog	1.0		themakeyourownzone	1.0
everydaylinuxuser	embassypages	1.0	**solmire**	embassypages	1.0
	preppyrunner	1.0		preppyrunner	1.0
	guidingtech	1.0		guidingtech	1.0
	nzhuntingandshooting	1.0		nzhuntingandshooting	1.0
	ftlauderdalewebcam	1.0		ftlauderdalewebcam	1.0
leisecamarica	invasaonerd	1.0	**sarahbakesgfree**	filmhafizasi	1.0
	caminhandojunto	1.0		sandraandwoo	1.0
	blogvambora	0.999		cleverlyinspired	1.0
	empreendedordigital	0.998		dcclothesline	1.0
	ligadoemserie	0.992		eastcoastcreativeblog	1.0
philosophersmag	embassypages	1.0	**najducokoliv**	rcmania	1.0
	preppyrunner	1.0		ireport	1.0
	guidingtech	1.0		novinky	1.0
	nzhuntingandshooting	1.0		securitymagazin	1.0
	ftlauderdalewebcam	1.0		prehled-prace	1.0
demoty	embassypages	1.0	**2nd-grade-math-salamanders**	embassypages	1.0
	preppyrunner	1.0		preppyrunner	1.0
	guidingtech	1.0		guidingtech	1.0
	nzhuntingandshooting	1.0		nzhuntingandshooting	1.0
	ftlauderdalewebcam	1.0		ftlauderdalewebcam	1.0
innatthecrossroads	patchworkposse	1.0	**aphroditewomenshealth**	embassypages	1.0
	filmhafizasi	1.0		preppyrunner	1.0
	sandraandwoo	1.0		guidingtech	1.0
	cleverlyinspired	1.0		nzhuntingandshooting	1.0
	dcclothesline	1.0		ftlauderdalewebcam	1.0
somethingswanky	filmhafizasi	1.0	**javaprogressivo**	midiaesportiva	1.0
	sandraandwoo	1.0		blogdohammer	1.0
	cleverlyinspired	1.0		mundodasmarcas	1.0
	dcclothesline	1.0		noitesinistra	0.999
	eastcoastcreativeblog	1.0		milicoponderao	0.999
queenbeecoupons	filmhafizasi	1.0	**mymotherlode**	filmhafizasi	1.0
	sandraandwoo	1.0		sandraandwoo	1.0
	cleverlyinspired	1.0		cleverlyinspired	1.0
	dcclothesline	1.0		dcclothesline	1.0
	eastcoastcreativeblog	1.0		eastcoastcreativeblog	1.0
irishsportsdaily	filmhafizasi	1.0	**forthefamily**	filmhafizasi	1.0
	sandraandwoo	1.0		sandraandwoo	1.0
	cleverlyinspired	1.0		cleverlyinspired	1.0
	dcclothesline	1.0		dcclothesline	1.0
	eastcoastcreativeblog	1.0		eastcoastcreativeblog	1.0
talkofweb	techydroid	1.0	**texastypeamom**	techydroid	1.0
	whoneedsmaps	1.0		whoneedsmaps	1.0
	westminsterkennelclub	1.0		westminsterkennelclub	1.0
	thisismyhappiness	1.0		thisismyhappiness	1.0
	themakeyourownzone	1.0		themakeyourownzone	1.0
mobilemela	embassypages	1.0	**beniarayan**	embassypages	1.0
	preppyrunner	1.0		preppyrunner	1.0
	guidingtech	1.0		guidingtech	1.0
	nzhuntingandshooting	1.0		nzhuntingandshooting	1.0
	ftlauderdalewebcam	1.0		ftlauderdalewebcam	1.0

Tabela E.3: Tabela Similaridade Base Canvas x 100 Consultas - Parte 3

Tabela Nível de Similaridade Base Canvas x 100 Consultas (Parte 3)					
Consulta	**Script Base Canvas**	**Similaridade**	**Consulta**	**Script Base Canvas**	**Similaridade**
jccnn	filmhafizasi	1.0	**upsda**	techydroid	1.0
	sandraandwoo	1.0		whoneedsmaps	1.0
	cleverlyinspired	1.0		westminsterkennelclub	1.0
	dcclothesline	1.0		thisismyhappiness	1.0
	eastcoastcreativeblog	1.0		themakeyourownzone	1.0
safa-ivrit	embassypages	1.0	**techsmart**	techydroid	1.0
	preppyrunner	1.0		whoneedsmaps	1.0
	guidingtech	1.0		westminsterkennelclub	1.0
	nzhuntingandshooting	1.0		thisismyhappiness	1.0
	ftlauderdalewebcam	1.0		themakeyourownzone	1.0
thecozyapron	techydroid	1.0	**adistanciaentre**	embassypages	1.0
	whoneedsmaps	1.0		preppyrunner	1.0
	westminsterkennelclub	1.0		guidingtech	1.0
	thisismyhappiness	1.0		nzhuntingandshooting	1.0
	themakeyourownzone	1.0		ftlauderdalewebcam	1.0
ricettedellanonna	embassypages	1.0	**bonbonbreak**	filmhafizasi	1.0
	preppyrunner	1.0		sandraandwoo	1.0
	guidingtech	1.0		cleverlyinspired	1.0
	nzhuntingandshooting	1.0		dcclothesline	1.0
	ftlauderdalewebcam	1.0		eastcoastcreativeblog	1.0
iscrapapp	filmhafizasi	1.0	**grammar**	embassypages	1.0
	sandraandwoo	1.0		preppyrunner	1.0
	cleverlyinspired	1.0		guidingtech	1.0
	dcclothesline	1.0		nzhuntingandshooting	1.0
	eastcoastcreativeblog	1.0		ftlauderdalewebcam	1.0
mobiany	osxarena	1.0	**zvartnots**	embassypages	1.0
	pixelshut	1.0		preppyrunner	1.0
	somethingtostream	1.0		guidingtech	1.0
	beanpanda	1.0		nzhuntingandshooting	1.0
	inati0n	1.0		ftlauderdalewebcam	1.0
robsessedpattinson	embassypages	1.0	**willcookforfriends**	filmhafizasi	1.0
	preppyrunner	1.0		sandraandwoo	1.0
	guidingtech	1.0		cleverlyinspired	1.0
	nzhuntingandshooting	1.0		dcclothesline	1.0
	ftlauderdalewebcam	1.0		eastcoastcreativeblog	1.0
thefirstmess	techydroid	1.0	**allmobiletools**	embassypages	1.0
	whoneedsmaps	1.0		preppyrunner	1.0
	westminsterkennelclub	1.0		guidingtech	1.0
	thisismyhappiness	1.0		nzhuntingandshooting	1.0
	themakeyourownzone	1.0		ftlauderdalewebcam	1.0
happyhealthymama	filmhafizasi	1.0	**wirelessadvisor**	embassypages	1.0
	sandraandwoo	1.0		preppyrunner	1.0
	cleverlyinspired	1.0		guidingtech	1.0
	dcclothesline	1.0		nzhuntingandshooting	1.0
	eastcoastcreativeblog	1.0		ftlauderdalewebcam	1.0
edutechupdates	embassypages	1.0	**beautythroughimperfection**	filmhafizasi	1.0
	preppyrunner	1.0		sandraandwoo	1.0
	guidingtech	1.0		cleverlyinspired	1.0
	nzhuntingandshooting	1.0		dcclothesline	1.0
	ftlauderdalewebcam	1.0		eastcoastcreativeblog	1.0
allmyvod	embassypages	1.0	**mynewestaddiction**	filmhafizasi	1.0
	preppyrunner	1.0		sandraandwoo	1.0
	guidingtech	1.0		cleverlyinspired	1.0
	nzhuntingandshooting	1.0		dcclothesline	1.0
	ftlauderdalewebcam	1.0		eastcoastcreativeblog	1.0
pattern-paradise	filmhafizasi	1.0	**littlemissmomma**	filmhafizasi	1.0
	sandraandwoo	1.0		sandraandwoo	1.0
	cleverlyinspired	1.0		cleverlyinspired	1.0
	dcclothesline	1.0		dcclothesline	1.0
	eastcoastcreativeblog	1.0		eastcoastcreativeblog	1.0
naivecookcooks	clarkscondensed	1.0	**lifeonvirginiastreet**	filmhafizasi	1.0
	11magnolialane	1.0		sandraandwoo	1.0
	kitchenstewardship	1.0		cleverlyinspired	1.0
	ihearteating	1.0		dcclothesline	1.0
	lexibites	1.0		eastcoastcreativeblog	1.0

Tabela E.4: Tabela Similaridade Base Canvas x 100 Consultas - Parte 4

Tabela Nível de Similaridade Base Canvas x 100 Consultas					
Consulta	**Script Base Canvas**	**Similaridade**	**Consulta**	**Script Base Canvas**	**Similaridade**
freecoin	mycoinads	1.0	**sugarbombed**	embassypages	1.0
	boxfaucet	1.0		preppyrunner	1.0
	bitcoinfaucetexchange	1.0		guidingtech	1.0
	freebtcfaucet	1.0		nzhuntingandshooting	1.0
	rekbitcoin	1.0		ftlauderdalewebcam	1.0
freewilliamsburg	filmhafizasi	1.0	**catalina-ponor**	filmhafizasi	1.0
	sandraandwoo	1.0		sandraandwoo	1.0
	cleverlyinspired	1.0		cleverlyinspired	1.0
	dcclothesline	1.0		dcclothesline	1.0
	eastcoastcreativeblog	1.0		eastcoastcreativeblog	1.0
natashalh	filmhafizasi	1.0	**yupitsvegan**	techydroid	1.0
	sandraandwoo	1.0		whoneedsmaps	1.0
	cleverlyinspired	1.0		westminsterkennelclub	1.0
	dcclothesline	1.0		thisismyhappiness	1.0
	eastcoastcreativeblog	1.0		themakeyourownzone	1.0
hackhackers	embassypages	1.0	**gooddinnermom**	filmhafizasi	1.0
	preppyrunner	1.0		sandraandwoo	1.0
	guidingtech	1.0		cleverlyinspired	1.0
	nzhuntingandshooting	1.0		dcclothesline	1.0
	ftlauderdalewebcam	1.0		eastcoastcreativeblog	1.0
theysmell	techydroid	1.0	**tankathon**	embassypages	1.0
	whoneedsmaps	1.0		preppyrunner	1.0
	westminsterkennelclub	1.0		guidingtech	1.0
	thisismyhappiness	1.0		nzhuntingandshooting	1.0
	themakeyourownzone	1.0		ftlauderdalewebcam	1.0
familyfuninomaha	techydroid	1.0	**uglyducklinghouse**	techydroid	1.0
	whoneedsmaps	1.0		whoneedsmaps	1.0
	westminsterkennelclub	1.0		westminsterkennelclub	1.0
	thisismyhappiness	1.0		thisismyhappiness	1.0
	themakeyourownzone	1.0		themakeyourownzone	1.0
thekitchenpaper	techydroid	1.0	**agentdunet**	embassypages	1.0
	whoneedsmaps	1.0		preppyrunner	1.0
	westminsterkennelclub	1.0		guidingtech	1.0
	thisismyhappiness	1.0		nzhuntingandshooting	1.0
	themakeyourownzone	1.0		ftlauderdalewebcam	1.0
feelingfoodish	techydroid	1.0	**foodies-magazin**	sanjorgeschool	1.0
	whoneedsmaps	1.0		laughspin	0.997
	westminsterkennelclub	1.0		viladoesporte	0.985
	thisismyhappiness	1.0		savingmoneylivingsmart	0.983
	themakeyourownzone	1.0		youknowyoulovefashion	0.983
prostoblog	embassypages	1.0	**littleblogonthehomestead**	itallstartedwithpaint	1.0
	preppyrunner	1.0		madetobeamomma	0.998
	guidingtech	1.0		azgrabaplate	0.997
	nzhuntingandshooting	1.0		theglamoroushousewife	0.997
	ftlauderdalewebcam	1.0		thewhitebuffalostylingco	0.997
urbangardensweb	techydroid	1.0	**news24zim**	outbreaknewstoday	1.0
	whoneedsmaps	1.0		take-me-to-auction	1.0
	westminsterkennelclub	1.0		urdunovels	0.999
	thisismyhappiness	1.0		shitlicious	0.999
	themakeyourownzone	1.0		freebiesjedi	0.999
quirkychrissy	kellyinthecity	1.0	**idgconnect**	reidl	0.987
	techranker	0.996		ch10	0.987
	splashmagazine	0.973		namebright	0.987
	veteranstoday	0.97		juwelo	0.987
	warriorsworld	0.97		piecesvoitures	0.986
opovonews	animexis	0.998	**globaltv**	manaserials	0.984
	willtirando	0.997		crisguerra	0.972
	lucrena	0.997		stream-a-ams1xx2sfcdnvideo5269	0.969
	criadoresdeconteudo	0.996		nsmb	0.96
	folhetimonline	0.996		hombremoderno	0.954
gerardofernandez	fotograpiya	0.998	**my-dictionary**	sa-mp	0.982
	dancemagazine	0.997		equinenow	0.979
	pointemagazine	0.997		thewisebaby	0.977
	sammichespsychmeds	0.997		honnest	0.973
	relyonhorror	0.996		videoblocks	0.968

Tabela E.5: Tabela Similaridade Base Canvas x 100 Consultas - Parte 5

Tabela Nível de Similaridade Base Canvas x 100 Consultas (Parte 5)					
Consulta	**Script Base**	**Similaridade**	**Consulta**	**Script Base Canvas**	**Similaridade**
koetube	jmunicipios	0.998	**2016election**	pistolsfiringblog	0.98
	animexis	0.998		akdirahost	0.954
	noticiasparaiba	0.997		tap-repeatedly	0.948
	zipix	0.997		fdlreporter	0.946
	criadoresdeconteudo	0.997		sheboyganpress	0.946
1057max	pistolsfiringblog	0.995	**neesoku**	matomame	0.967
	bradfordcityfc	0.939		point-island	0.967
	destinydb	0.933		soccer-douga	0.965
	typingtestnow	0.931		nakluky	0.765
	myschool	0.931		platinumblackcomic	0.745
spreaker	concertpass	0.99	**guesstheemoji-answers**	booshsports	0.936
	eventim	0.99		tohapi	0.921
	a1-forum	0.99		date	0.891
	ticketcorner	0.99		freeonlinephotoeditor	0.884
	domosfera	0.989		fastsocialfollower	0.868
southendnewsnetwork	theurbantwist	0.987	**mobilinkmobile**	staffordshirenewsletter	0.778
	im-internetmarketing	0.987		grantspassweather	0.776
	donandroid	0.987		sigalert	0.767
	devzum	0.987		tamilmagazines	0.742
	trulygeeky	0.987		mixanitouxronou	0.701

Apêndice F

Similaridade Top 5 Phishtank

Tabela F.1: Tabela Similaridade Base Phishtank x 100 Consultas - Parte 1

Tabela Nível de Similaridade Base Phishtank x 100 Consultas (Parte 1)					
Consulta	**Script Base Phishtank**	**Similaridade**	**Consulta**	**Script Base Phishtank**	**Similaridade**
themamamaven	caraudioacapulco	1.0	**wespecify**	caraudioacapulco	1.0
	pracadarepublicaembeja	1.0		pracadarepublicaembeja	1.0
	sigarabirakmak	1.0		sigarabirakmak	1.0
	pasarpedia	1.0		pasarpedia	1.0
	jimmyseas	1.0		jimmyseas	1.0
hoboken411	sigarabirakmak	1.0	**ongoingpro**	sigarabirakmak	1.0
	replacementroofingtx	1.0		replacementroofingtx	1.0
	caraudioacapulco	1.0		caraudioacapulco	1.0
	davinciresidence	1.0		davinciresidence	1.0
	rcacas	1.0		rcacas	1.0
meridianstar	maisponto	1.0	**everydaylinuxuser**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0
openairfarmersmarkets	maisponto	1.0	**solmire**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0
thediyvillage	caraudioacapulco	1.0	**philosophersmag**	maisponto	1.0
	pracadarepublicaembeja	1.0		infobel	1.0
	sigarabirakmak	1.0		ricardoeletro2	1.0
	pasarpedia	1.0		info-setting2016	1.0
	jimmyseas	1.0		lagosstatenews	1.0
powerfulmothering	sigarabirakmak	1.0	**demoty**	maisponto	1.0
	replacementroofingtx	1.0		infobel	1.0
	caraudioacapulco	1.0		ricardoeletro2	1.0
	davinciresidence	1.0		info-setting2016	1.0
	rcacas	1.0		lagosstatenews	1.0
dokka	maisponto	1.0	**innatthecrossroads**	jimmyseas	1.0
	infobel	1.0		replacementroofingtx	1.0
	ricardoeletro2	1.0		caraudioacapulco	1.0
	info-setting2016	1.0		davinciresidence	1.0
	lagosstatenews	1.0		rcacas	1.0

Tabela F.2: Tabela Similaridade Base Phishtank x 100 Consultas - Parte 2

Tabela Nível de Similaridade Base Phishtank x 100 Consultas (Parte 2)					
Consulta	**Script Base Phishtank**	**Similaridade**	**Consulta**	**Script Base Phishtank**	**Similaridade**
wunderwuensche	maisponto	1.0	**somethingswanky**	sigarabirakmak	1.0
	infobel	1.0		replacementroofingtx	1.0
	ricardoeletro2	1.0		caraudioacapulco	1.0
	info-setting2016	1.0		davinciresidence	1.0
	lagosstatenews	1.0		rcacas	1.0
artilhariadigital	maisponto	1.0	**queenbeecoupons**	sigarabirakmak	1.0
	infobel	1.0		replacementroofingtx	1.0
	ricardoeletro2	1.0		caraudioacapulco	1.0
	info-setting2016	1.0		davinciresidence	1.0
	lagosstatenews	1.0		rcacas	1.0
the-socialites-closet	caraudioacapulco	1.0	**irishsportsdaily**	sigarabirakmak	1.0
	pracadarepublicaembeja	1.0		replacementroofingtx	1.0
	sigarabirakmak	1.0		caraudioacapulco	1.0
	pasarpedia	1.0		davinciresidence	1.0
	jimmyseas	1.0		rcacas	1.0
eodisha	maisponto	1.0	**talkofweb**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
mobilemela	maisponto	1.0	**forthefamily**	sigarabirakmak	1.0
	infobel	1.0		replacementroofingtx	1.0
	ricardoeletro2	1.0		caraudioacapulco	1.0
	info-setting2016	1.0		davinciresidence	1.0
	lagosstatenews	1.0		rcacas	1.0
jccnn	sigarabirakmak	1.0	**texastypeamom**	caraudioacapulco	1.0
	replacementroofingtx	1.0		pracadarepublicaembeja	1.0
	caraudioacapulco	1.0		sigarabirakmak	1.0
	davinciresidence	1.0		pasarpedia	1.0
	rcacas	1.0		jimmyseas	1.0
safa-ivrit	maisponto	1.0	**beniarayan**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0
sarahbakesgfree	sigarabirakmak	1.0	**upsda**	caraudioacapulco	1.0
	replacementroofingtx	1.0		pracadarepublicaembeja	1.0
	caraudioacapulco	1.0		sigarabirakmak	1.0
	davinciresidence	1.0		pasarpedia	1.0
	rcacas	1.0		jimmyseas	1.0
2nd-grade-math-salamanders	maisponto	1.0	**techsmart**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
aphroditewomenshealth	maisponto	1.0	**thefirstmess**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
mymotherlode	sigarabirakmak	1.0	**happyhealthymama**	sigarabirakmak	1.0
	replacementroofingtx	1.0		replacementroofingtx	1.0
	caraudioacapulco	1.0		caraudioacapulco	1.0
	davinciresidence	1.0		davinciresidence	1.0
	rcacas	1.0		rcacas	1.0
thecozyapron	caraudioacapulco	1.0	**edutechupdates**	maisponto	1.0
	pracadarepublicaembeja	1.0		infobel	1.0
	sigarabirakmak	1.0		ricardoeletro2	1.0
	pasarpedia	1.0		info-setting2016	1.0
	jimmyseas	1.0		lagosstatenews	1.0
ricettedellanonna	maisponto	1.0	**allmyvod**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0

Tabela F.3: Tabela Similaridade Base Phishtank x 100 Consultas - Parte 3

Tabela Nível de Similaridade Base Phishtank x 100 Consultas (Parte 3)					
Consulta	**Script Base Phishtank**	**Similaridade**	**Consulta**	**Script Base Phishtank**	**Similaridade**
iscrapapp	sigarabirakmak	1.0	**pattern-paradise**	sigarabirakmak	1.0
	replacementroofingtx	1.0		replacementroofingtx	1.0
	caraudioacapulco	1.0		caraudioacapulco	1.0
	davinciresidence	1.0		davinciresidence	1.0
	rcacas	1.0		rcacas	1.0
robsessedpattinson	maisponto	1.0	**adistanciaentre**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0
bonbonbreak	sigarabirakmak	1.0	**natashalh**	sigarabirakmak	1.0
	replacementroofingtx	1.0		replacementroofingtx	1.0
	caraudioacapulco	1.0		caraudioacapulco	1.0
	davinciresidence	1.0		davinciresidence	1.0
	rcacas	1.0		rcacas	1.0
grammar	maisponto	1.0	**hackhackers**	maisponto	1.0
	infobel	1.0		infobel	1.0
	ricardoeletro2	1.0		ricardoeletro2	1.0
	info-setting2016	1.0		info-setting2016	1.0
	lagosstatenews	1.0		lagosstatenews	1.0
zvartnots	maisponto	1.0	**theysmell**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
willcookforfriends	sigarabirakmak	1.0	**familyfuninomaha**	caraudioacapulco	1.0
	replacementroofingtx	1.0		pracadarepublicaembeja	1.0
	caraudioacapulco	1.0		sigarabirakmak	1.0
	davinciresidence	1.0		pasarpedia	1.0
	rcacas	1.0		jimmyseas	1.0
allmobiletools	maisponto	1.0	**thekitchenpaper**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
wirelessadvisor	maisponto	1.0	**feelingfoodish**	caraudioacapulco	1.0
	infobel	1.0		pracadarepublicaembeja	1.0
	ricardoeletro2	1.0		sigarabirakmak	1.0
	info-setting2016	1.0		pasarpedia	1.0
	lagosstatenews	1.0		jimmyseas	1.0
beautythroughimperfection	sigarabirakmak	1.0	**prostoblog**	maisponto	1.0
	replacementroofingtx	1.0		infobel	1.0
	caraudioacapulco	1.0		ricardoeletro2	1.0
	davinciresidence	1.0		info-setting2016	1.0
	rcacas	1.0		lagosstatenews	1.0
mynewestaddiction	sigarabirakmak	1.0	**urbangardensweb**	caraudioacapulco	1.0
	replacementroofingtx	1.0		pracadarepublicaembeja	1.0
	caraudioacapulco	1.0		sigarabirakmak	1.0
	davinciresidence	1.0		pasarpedia	1.0
	rcacas	1.0		jimmyseas	1.0
littlemissmomma	sigarabirakmak	1.0	**sugarbombed**	maisponto	1.0
	replacementroofingtx	1.0		infobel	1.0
	caraudioacapulco	1.0		ricardoeletro2	1.0
	davinciresidence	1.0		info-setting2016	1.0
	rcacas	1.0		lagosstatenews	1.0
lifeonvirginiastreet	sigarabirakmak	1.0	**catalina-ponor**	sigarabirakmak	1.0
	replacementroofingtx	1.0		replacementroofingtx	1.0
	caraudioacapulco	1.0		caraudioacapulco	1.0
	davinciresidence	1.0		davinciresidence	1.0
	rcacas	1.0		rcacas	1.0
freewilliamsburg	sigarabirakmak	1.0	**yupitsvegan**	caraudioacapulco	1.0
	replacementroofingtx	1.0		pracadarepublicaembeja	1.0
	caraudioacapulco	1.0		sigarabirakmak	1.0
	davinciresidence	1.0		pasarpedia	1.0
	rcacas	1.0		jimmyseas	1.0

Tabela F.4: Tabela Similaridade Base Phishtank x 100 Consultas - Parte 4

Tabela Nível de Similaridade Base Phishtank x 100 Consultas (Parte 4)					
Consulta	**Script Base Phishtank**	**Similaridade**	**Consulta**	**Script Base Phishtank**	**Similaridade**
gooddinnermom	sigarabirakmak	1.0	**southendnewsnetwork**	ar	0.992
	replacementroofingtx	1.0		myhotmailsignin	0.981
	caraudioacapulco	1.0		beffy	0.981
	davinciresidence	1.0		correodegmail	0.981
	rcacas	1.0		secure-banklogin	0.976
tankathon	maisponto	1.0	**javaprogressivo**	sa	0.989
	infobel	1.0		bnpparibas-abstract-expressionism	0.797
	ricardoeletro2	1.0		mali-dugi	0.797
	info-setting2016	1.0		greenhub	0.766
	lagosstatenews	1.0		becgiemob	0.68
uglyducklinghouse	caraudioacapulco	1.0	**news24eg**	pallascaaldia	0.988
	pracadarepublicaembeja	1.0		fanatiksport	0.965
	sigarabirakmak	1.0		kitchensrus	0.956
	pasarpedia	1.0		eclecticgrape	0.954
	jimmyseas	1.0		lioninthesun	0.954
agentdunet	maisponto	1.0	**songbirdgarden**	sa	0.987
	infobel	1.0		greenhub	0.779
	ricardoeletro2	1.0		bnpparibas-abstract-expressionism	0.766
	info-setting2016	1.0		mali-dugi	0.766
	lagosstatenews	1.0		promolinks	0.672
epicobottles	altavistawines	1.0	**news24zim**	pallascaaldia	0.983
	es	1.0		kitchensrus	0.974
	yonearts	1.0		lioninthesun	0.972
	tra	1.0		hoop4eva	0.972
	specialchild	0.999		jamiexxperth	0.972
naivecookcooks	syengage	1.0	**littleblogonthehomestead**	goo	0.981
	pejdah-pharmacia	1.0		santechno	0.969
	ibooking	0.996		mashavrelocation	0.969
	buildonlinewealth	0.993		kitchensrus	0.969
	clermontlounge	0.992		beast-the-animal	0.969
mobiany	loginsign	0.999	**opovonews**	sa	0.981
	aliexpress	0.999		greenhub	0.792
	secure-banklogin	0.998		bnpparibas-abstract-expressionism	0.773
	giftcardblogger	0.992		dugi	0.773
	banksignin	0.988		promolinks	0.747
androidtv	moneyexprt	0.998	**gerardofernandez**	lcmaquinasseladoras	0.978
	familiatuccini	0.998		serrauquer	0.973
	unitygamesbox	0.997		jehansen	0.972
	nubeviajera	0.981		healthinsurance	0.971
	pousadacavalinno	0.981		eddaturkey	0.971
foodies-magazin	hism	0.997	**spreaker**	courchevel	0.975
	notafog2016	0.993		worldsalsachampionships	0.968
	mbu	0.989		kazichschool	0.963
	911beautystudios	0.983		theworkouts	0.963
	choicesunlimited	0.983		machizo	0.962
jqueryhouse	courchevel	0.995	**koetube**	sa	0.972
	worldsalsachampionships	0.989		greenhub	0.757
	theworkouts	0.986		promolinks	0.756
	machizo	0.986		bnpparibas-abstract-expressionism	0.737
	nadiabernocchi	0.983		mali-dugi	0.737
quirkychrissy	syengage	0.993	**freecoin**	greenhub	0.972
	pejdah-pharmacia	0.993		twinyoubethinking	0.871
	ibooking	0.989		becgiemob	0.852
	dianagarces	0.988		ipkill	0.828
	clermontlounge	0.987		clean-clean-peru	0.804
2016election	en	0.967	**globaltv**	sa	0.904
	progorod	0.962		bnpparibas-abstract-expressionism	0.858
	forexmoneyback	0.947		mali-dugi	0.858
	ba-tango	0.945		mybeltz	0.775
	batallonchacras	0.943		greenhub	0.762
idgconnect	lilicoimoveis	0.953	**leisecamarica**	sa	0.902
	spotcampus	0.953		promolinks	0.855
	hyperbarichealingcenter	0.907		evenpro	0.781
	footspecialistbrampton	0.893		nolagroup	0.761
	innmoema	0.893		flatbellyfitness	0.76

Tabela F.5: Tabela Similaridade Base Phishtank x 100 Consultas - Parte 5

Tabela Nível de Similaridade Base Phishtank x Consultas (Parte 5)					
Consulta Canvas	**Script Base Phishtank**	**Similaridade**	**Consulta**	**Script Base Phishtank**	**Similaridade**
1057max	en	0.942	**neesoku**	esivah	0.836
	progorod	0.937		elogix	0.586
	novatekit	0.923		sire-china	0.569
	ibusinessolution	0.92		carverslaw	0.567
	ba-tango	0.919		hackspirit	0.56
my-dictionary	ibusinessolution	0.929	**havadurumux**	create-new-account	0.827
	biovene	0.927		t-alquds	0.824
	stevewarshak	0.921		farlainlake	0.819
	wistub-brenner	0.918		keeneteamvegas	0.797
	capalaroche	0.917		ar	0.766
guesstheemoji-answers	eurochistka	0.923	**mobilinkmobile**	danielabrantes	0.689
	itpbacau	0.865		capitalcu	0.67
	globalcleaning	0.836		dhaainkanbaa	0.648
	wisefellas	0.816		dotartprinting	0.638
	wibs	0.81		autoscriba	0.631
roadglide	dotartprinting	0.921	**najducokoliv**	linkedin	0.596
	couponshe	0.861		plasticmonkey	0.596
	ronarhost	0.847		sportyfit	0.596
	promowear	0.847		greenhub	0.551
	bancofotografias	0.838		bnpparibas-abstract-expressionism	0.523

Apêndice G

Similaridade Top 5 DMOZ

Tabela G.1: Tabela Similaridade Base Dmoz x 100 Consultas - Parte 1

Tabela Nível de Similaridade Base Dmoz x 100 Consultas (Parte 1)					
Consulta	**Script Base Dmoz**	**Similaridade**	**Consulta**	**Script Base Dmoz**	**Similaridade**
meridianstar	shesmoke	1.0	**safa-ivrit**	highpowergraphics	1.0
	plum	1.0		shesmoke	1.0
	bobthealien	1.0		plum	1.0
	cottonclouds	1.0		bobthealien	1.0
	madmeatgenius	1.0		cottonclouds	1.0
openairfarmersmarkets	highpowergraphics	1.0	**2nd-grade-math-salamanders**	highpowergraphics	1.0
	shesmoke	1.0		shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
dokka	highpowergraphics	1.0	**aphroditewomenshealth**	highpowergraphics	1.0
	shesmoke	1.0		shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
wunderwuensche	highpowergraphics	1.0	**ricettedellanonna**	highpowergraphics	1.0
	shesmoke	1.0		shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
artilhariadigital	shesmoke	1.0	**beniarayan**	shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
	madmeatgenius	1.0		madmeatgenius	1.0
eodisha	shesmoke	1.0	**robsessedpattinson**	shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
	madmeatgenius	1.0		madmeatgenius	1.0
everydaylinuxuser	highpowergraphics	1.0	**edutechupdates**	shesmoke	1.0
	shesmoke	1.0		plum	1.0
	plum	1.0		bobthealien	1.0
	bobthealien	1.0		cottonclouds	1.0
	cottonclouds	1.0		madmeatgenius	1.0

Tabela G.2: Tabela Similaridade Base Dmoz x 100 Consultas - Parte 2

Tabela Nível de Similaridade Base Dmoz x 100 Consultas (Parte 2)					
Consulta	**Script Base DMOZ**	**Similaridade**	**Consulta**	**Script Base DMOZ**	**Similaridade**
solmire	shesmoke	1.0	**allmyvod**	highpowergraphics	1.0
	plum	1.0		shesmoke	1.0
	bobthealien	1.0		plum	1.0
	cottonclouds	1.0		bobthealien	1.0
	madmeatgenius	1.0		cottonclouds	1.0
philosophersmag	shesmoke	1.0	**adistanciaentre**	shesmoke	1.0
	plum	1.0		plum	1.0
	bobthealien	1.0		bobthealien	1.0
	cottonclouds	1.0		cottonclouds	1.0
	madmeatgenius	1.0		madmeatgenius	1.0
demoty	highpowergraphics	1.0	**grammar**	shesmoke	1.0
	shesmoke	1.0		plum	1.0
	plum	1.0		bobthealien	1.0
	bobthealien	1.0		cottonclouds	1.0
	cottonclouds	1.0		madmeatgenius	1.0
mobilemela	highpowergraphics	1.0	**zvartnots**	shesmoke	1.0
	shesmoke	1.0		plum	1.0
	plum	1.0		bobthealien	1.0
	bobthealien	1.0		cottonclouds	1.0
	cottonclouds	1.0		madmeatgenius	1.0
allmobiletools	shesmoke	1.0	**freecoin**	networkworld	0.958
	plum	1.0		itworld	0.958
	bobthealien	1.0		23d	0.955
	cottonclouds	1.0		aidsinfo	0.805
	madmeatgenius	1.0		herrschners	0.718
wirelessadvisor	highpowergraphics	1.0	**spreaker**	26and2	0.958
	shesmoke	1.0		ker	0.958
	plum	1.0		earthdance	0.957
	bobthealien	1.0		flairstrips	0.957
	cottonclouds	1.0		goalballuk	0.957
hackhackers	shesmoke	1.0	**gerardofernandez**	bhavayoga	0.957
	plum	1.0		gaetanomansi	0.957
	bobthealien	1.0		robinsonarchive	0.955
	cottonclouds	1.0		ellenbogen	0.955
	madmeatgenius	1.0		productioncraft	0.953
prostoblog	shesmoke	1.0	**mobiany**	plausiblefutures	0.948
	plum	1.0		howtobbqright	0.946
	bobthealien	1.0		leaguefreak	0.942
	cottonclouds	1.0		accessgenealogy	0.942
	madmeatgenius	1.0		cbrdigital	0.939
sugarbombed	shesmoke	1.0	**news24zim**	pancero	0.944
	plum	1.0		jnproductions	0.944
	bobthealien	1.0		meat	0.939
	cottonclouds	1.0		ontariogenomics	0.924
	madmeatgenius	1.0		atmananda	0.918
tankathon	shesmoke	1.0	**littleblogonthehomestead**	kingsford	0.941
	plum	1.0		jnproductions	0.939
	bobthealien	1.0		pancero	0.934
	cottonclouds	1.0		kamadojim	0.929
	madmeatgenius	1.0		simonsmd	0.927
agentdunet	highpowergraphics	1.0	**1057max**	accentonline	0.942
	shesmoke	1.0		petmd	0.94
	plum	1.0		prenatalyogacenter	0.934
	bobthealien	1.0		onlinelibrary	0.904
	cottonclouds	1.0		abcnews	0.875
androidtv	cosmoquest	1.0	**2016election**	accentonline	0.94
	fezana	0.999		prenatalyogacenter	0.94
	homebbq	0.999		petmd	0.936
	biggreeneggic	0.999		onlinelibrary	0.921
	grillgirl	0.999		ams	0.899
jqueryhouse	26and2	0.981	**epicobottles**	eplbas	0.939
	ker	0.981		robotshop	0.939
	earthdance	0.98		cyndilee	0.937
	flairstrips	0.98		stillpointyogastudios	0.937
	goalballuk	0.98		accompanyvideo	0.937

Tabela G.3: Tabela Similaridade Base Dmoz x 100 Consultas - Parte 3

Tabela Nível de Similaridade Base Dmoz x 100 Consultas (Parte 3)					
Consulta	**Script Base DMOZ**	**Similaridade**	**Consulta**	**Script Base DMOZ**	**Similaridade**
foodies-magazin	earthdance	0.98	**themamamaven**	mdsafrica	0.933
	flairstrips	0.98		mypet-memorial	0.933
	goalballuk	0.98		sonicyoga	0.933
	26and2	0.979		modern-rocket	0.933
	ker	0.979		panta-rhei	0.933
my-dictionary	abcnews	0.975	**hoboken411**	mdsafrica	0.933
	prenatalyogacenter	0.93		mypet-memorial	0.933
	onlinelibrary	0.925		sonicyoga	0.933
	accentonline	0.917		modern-rocket	0.933
	petmd	0.908		panta-rhei	0.933
innatthecrossroads	mdsafrica	0.939	**mymotherlode**	mdsafrica	0.933
	mypet-memorial	0.939		mypet-memorial	0.933
	sonicyoga	0.939		sonicyoga	0.933
	modern-rocket	0.939		modern-rocket	0.933
	panta-rhei	0.939		panta-rhei	0.933
the-socialites-closet	mdsafrica	0.933	**thecozyapron**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
wespecify	mdsafrica	0.933	**iscrapapp**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
powerfulmothering	mdsafrica	0.933	**forthefamily**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
ongoingpro	mdsafrica	0.933	**texastypeamom**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
somethingswanky	mdsafrica	0.933	**upsda**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
queenbeecoupons	mdsafrica	0.933	**techsmart**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
irishsportsdaily	mdsafrica	0.933	**thefirstmess**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
talkofweb	mdsafrica	0.933	**happyhealthymama**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
jccnn	mdsafrica	0.933	**pattern-paradise**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
sarahbakesgfree	mdsafrica	0.933	**bonbonbreak**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933

Tabela G.4: Tabela Similaridade Base Dmoz x 100 Consultas - Parte 4

Tabela Nível de Similaridade Base Dmoz x 100 Consultas (Parte 4)					
Consulta	**Script Base DMOZ**	**Similaridade**	**Consulta**	**Script Base DMOZ**	**Similaridade**
willcookforfriends	mdsafrica	0.933	**thekitchenpaper**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
beautythroughimperfection	mdsafrica	0.933	**feelingfoodish**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
mynewestaddiction	mdsafrica	0.933	**urbangardensweb**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
littlemissmomma	mdsafrica	0.933	**catalina-ponor**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
lifeonvirginiastreet	mdsafrica	0.933	**yupitsvegan**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
freewilliamsburg	mdsafrica	0.933	**gooddinnermom**	mdsafrica	0.933
	mypet-memorial	0.933		mypet-memorial	0.933
	sonicyoga	0.933		sonicyoga	0.933
	modern-rocket	0.933		modern-rocket	0.933
	panta-rhei	0.933		panta-rhei	0.933
natashalh	mdsafrica	0.933	**naivecookcooks**	mdsafrica	0.928
	mypet-memorial	0.933		mypet-memorial	0.928
	sonicyoga	0.933		sonicyoga	0.928
	modern-rocket	0.933		modern-rocket	0.928
	panta-rhei	0.933		panta-rhei	0.928
theysmell	mdsafrica	0.933	**news24eg**	pancero	0.928
	mypet-memorial	0.933		jnproductions	0.928
	sonicyoga	0.933		smoking-meat	0.923
	modern-rocket	0.933		ontariogenomics	0.908
	panta-rhei	0.933		atmananda	0.902
uglyducklinghouse	mdsafrica	0.933	**quirkychrissy**	richiegillphotography	0.926
	mypet-memorial	0.933		lifeinmotion	0.924
	sonicyoga	0.933		beyondtrust	0.924
	modern-rocket	0.933		mypet-memorial	0.922
	panta-rhei	0.933		safekids	0.922
thediyvillage	mdsafrica	0.933	**idgconnect**	cambo	0.918
	mypet-memorial	0.933		doublezoot	0.864
	sonicyoga	0.933		manningdigital	0.863
	modern-rocket	0.933		kingsford	0.863
	panta-rhei	0.933		simonsmd	0.862
familyfuninomaha	mdsafrica	0.933	**southendnewsnetwork**	cbrdigital	0.917
	mypet-memorial	0.933		bbqsmokersite	0.916
	sonicyoga	0.933		petlossjourney	0.916
	modern-rocket	0.933		stopthefleas	0.916
	panta-rhei	0.933		plausiblefutures	0.914
guesstheemoji-answers	asr-gooyesh	0.903	**leisecamarica**	himalayaninstitute	0.733
	horizonvp	0.9		mdsafrica	0.732
	advmediaservices	0.87		mypet-memorial	0.732
	kodak	0.869		sonicyoga	0.732
	kidoz	0.868		modern-rocket	0.732
globaltv	renewingthecountryside	0.889	**neesoku**	houstonyoga	0.723
	staylor-made	0.716		apartment-ideas	0.698
	23d	0.708		fibergarden	0.697
	softcat	0.69		eplbas	0.695
	ccvideo	0.682		countrysidenetwork	0.693

Tabela G.5: Tabela Similaridade Base Dmoz x 100 Consultas - Parte 5

Tabela Nível de Similaridade Base Dmoz x 100 Consultas (Parte 5)					
Consulta	**Script Base Dmoz**	**Similaridade**	**Consulta**	**Script Base Dmoz**	**Similaridade**
havadurumux	dellarte	0.875	**javaprogressivo**	23d	0.732
	rjmuna	0.873		renewingthecountryside	0.712
	andreschuster	0.869		networkworld	0.696
	apartment-ideas	0.867		itworld	0.696
	countrysidenetwork	0.834		ccvideo	0.65
roadglide	cowgirlscountry	0.873	**koetube**	23d	0.72
	barbequemaster	0.873		networkworld	0.702
	astronomicaloptical	0.873		itworld	0.702
	farewellfurryfriend	0.873		paintballtraining	0.673
	cloudbyuchit	0.873		himalayaninstitute	0.668
opovonews	23d	0.754	**mobilinkmobile**	farewellfurryfriend	0.715
	networkworld	0.749		cloudbyuchit	0.715
	itworld	0.749		grillingandsmoking	0.715
	paintballtraining	0.691		cowgirlscountry	0.715
	himalayaninstitute	0.685		barbequemaster	0.715
songbirdgarden	23d	0.739	**najducokoliv**	23d	0.616
	networkworld	0.677		networkworld	0.603
	itworld	0.677		itworld	0.603
	renewingthecountryside	0.662		aidsinfo	0.558
	aidsinfo	0.624		renewingthecountryside	0.554

Apêndice H

Similaridade Top 5 Alexa

Tabela H.1: Tabela Similaridade Base Alexa x 100 Consultas - Parte 1

Tabela Nível de Similaridade Base Alexa x 100 Consultas (Parte 1)					
Consulta	**Script Base Alexa**	**Similaridade**	**Consulta**	**Script Base Alexa**	**Similaridade**
meridianstar	bussolaescolar	1.0	**safa-ivrit**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
openairfarmersmarkets	bussolaescolar	1.0	**2nd-grade-math-salamanders**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
dokka	bussolaescolar	1.0	**aphroditewomenshealth**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
wunderwuensche	bussolaescolar	1.0	**ricettedellanonna**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
artilhariadigital	bussolaescolar	1.0	**beniarayan**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
eodisha	bussolaescolar	1.0	**robsessedpattinson**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
everydaylinuxuser	bussolaescolar	1.0	**edutechupdates**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0

Tabela H.2: Tabela Similaridade Base Alexa x 100 Consultas - Parte 2

Tabela Nível de Similaridade Base Alexa x 100 Consultas (Parte 2)					
Consulta	**Script Base Alexa**	**Similaridade**	**Consulta**	**Script Base Alexa**	**Similaridade**
solmire	bussolaescolar	1.0	**allmyvod**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
philosophersmag	bussolaescolar	1.0	**adistanciaentre**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
demoty	bussolaescolar	1.0	**grammar**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
mobilemela	bussolaescolar	1.0	**zvartnots**	bussolaescolar	1.0
	cidade-brasil	1.0		cidade-brasil	1.0
	amofilmeshd	1.0		amofilmeshd	1.0
	batepapo	1.0		batepapo	1.0
	biomedicinapadrao	1.0		biomedicinapadrao	1.0
allmobiletools	bussolaescolar	1.0	**globaltv**	album	0.995
	cidade-brasil	1.0		clickjogos	0.964
	amofilmeshd	1.0		blogdouo	0.89
	batepapo	1.0		clicfolha	0.888
	biomedicinapadrao	1.0		avph	0.887
wirelessadvisor	bussolaescolar	1.0	**jqueryhouse**	aquasn	0.995
	cidade-brasil	1.0		citisystems	0.984
	amofilmeshd	1.0		academiadomarketing	0.984
	batepapo	1.0		cabralsilvaadvogados	0.984
	biomedicinapadrao	1.0		10i9	0.984
hackhackers	bussolaescolar	1.0	**koetube**	ahzeus	0.992
	cidade-brasil	1.0		blogdoprimo	0.99
	amofilmeshd	1.0		animexis	0.99
	batepapo	1.0		blogvambora	0.983
	biomedicinapadrao	1.0		clubedovectra	0.983
prostoblog	bussolaescolar	1.0	**freecoin**	animaniaclub	0.992
	cidade-brasil	1.0		atoananet	0.992
	amofilmeshd	1.0		cinema10	0.992
	batepapo	1.0		belezanaweb	0.991
	biomedicinapadrao	1.0		classificados	0.984
sugarbombed	bussolaescolar	1.0	**foodies-magazin**	academiadomarketing	0.986
	cidade-brasil	1.0		cabralsilvaadvogados	0.986
	amofilmeshd	1.0		10i9	0.986
	batepapo	1.0		0800net	0.986
	biomedicinapadrao	1.0		citisystems	0.984
tankathon	bussolaescolar	1.0	**spreaker**	aquasn	0.984
	cidade-brasil	1.0		citisystems	0.972
	amofilmeshd	1.0		academiadomarketing	0.971
	batepapo	1.0		cabralsilvaadvogados	0.971
	biomedicinapadrao	1.0		10i9	0.971
agentdunet	bussolaescolar	1.0	**gerardofernandez**	clinicastop10	0.982
	cidade-brasil	1.0		cittati	0.982
	amofilmeshd	1.0		baboo	0.978
	batepapo	1.0		autopolis	0.977
	biomedicinapadrao	1.0		artritereumatoide	0.977
javaprogressivo	actionsecomics2	0.999	**songbirdgarden**	clicfolha	0.98
	avph	0.997		blogdouo	0.98
	blogluhfernandez	0.997		avph	0.98
	areavip	0.997		blogluhfernandez	0.98
	blogvambora	0.997		animexis	0.979
androidtv	clickcamboriu	0.999	**idgconnect**	achecep	0.979
	alemdaruaatelier	0.999		atrevida	0.941
	agoramt	0.999		bagaggio	0.941
	andrezadicaeindicadisney	0.999		ceciliadale	0.941
	1001cupomdedescontos	0.999		afabula	0.939

Tabela H.3: Tabela Similaridade Base Alexa x 100 Consultas - Parte 3

Tabela Nível de Similaridade Base Alexa x 100 Consultas (Parte 3)					
Consulta	**Script Base Alexa**	**Similaridade**	**Consulta**	**Script Base Alexa**	**Similaridade**
leisecamarica	atentados	0.998	**my-dictionary**	agu	0.976
	chato	0.998		amcham	0.975
	acessaber	0.991		brascomm	0.957
	agendapesquisa	0.991		belasaude	0.954
	campos24horas	0.978		centraldemarcacao	0.953
opovonews	ahzeus	0.997	**news24eg**	antiinflamatorios	0.951
	animexis	0.995		apelidos	0.951
	blogdoprimo	0.991		anticoncepcionais	0.946
	avph	0.987		bheliodora	0.945
	blogluhfernandez	0.987		bloginformaticamicrocamp	0.942
2016election	cdlrio	0.959	**hoboken411**	caadf	0.874
	acessoainformacao	0.954		agorams	0.874
	cfa	0.954		bundasgostosas	0.874
	bloggerenciado	0.953		aiesec	0.874
	brascomm	0.948		bigshopping	0.874
news24zim	antiinflamatorios	0.956	**thediyvillage**	caadf	0.874
	apelidos	0.956		agorams	0.874
	bheliodora	0.953		bundasgostosas	0.874
	anticoncepcionais	0.952		aiesec	0.874
	bloginformaticamicrocamp	0.95		bigshopping	0.874
guesstheemoji-answers	amominhacelula	0.948	**powerfulmothering**	caadf	0.874
	c-date	0.857		agorams	0.874
	cambiodeboys	0.847		bundasgostosas	0.874
	blogauto	0.842		aiesec	0.874
	blogdofernandomesquita	0.84		bigshopping	0.874
1057max	cdlrio	0.947	**the-socialites-closet**	caadf	0.874
	acessoainformacao	0.946		agorams	0.874
	cfa	0.946		bundasgostosas	0.874
	brascomm	0.945		aiesec	0.874
	canalconecta	0.943		bigshopping	0.874
littleblogonthehomestead	bloginformaticamicrocamp	0.92	**wespecify**	caadf	0.874
	antiinflamatorios	0.919		agorams	0.874
	apelidos	0.919		bundasgostosas	0.874
	bheliodora	0.917		aiesec	0.874
	autoracing	0.917		bigshopping	0.874
roadglide	adctec	0.904	**ongoingpro**	caadf	0.874
	bettbrasileducar	0.851		agorams	0.874
	abdi	0.843		bundasgostosas	0.874
	arazao	0.843		aiesec	0.874
	androidlista	0.836		bigshopping	0.874
mobiany	blogdoemprego	0.896	**somethingswanky**	caadf	0.874
	baixarwhatsapp	0.896		agorams	0.874
	adorosaude	0.893		bundasgostosas	0.874
	brasileirosporbuenosaires	0.893		aiesec	0.874
	blogdopaz	0.891		bigshopping	0.874
southendnewsnetwork	acezone	0.888	**queenbeecoupons**	caadf	0.874
	clubdofitness	0.887		agorams	0.874
	ahduvido	0.877		bundasgostosas	0.874
	blogtantofaz	0.876		aiesec	0.874
	blogdonelio	0.87		bigshopping	0.874
epicobottles	bluebus	0.887	**irishsportsdaily**	caadf	0.874
	alafiadigital	0.884		agorams	0.874
	amorc	0.883		bundasgostosas	0.874
	bandeiradois	0.881		aiesec	0.874
	bhi	0.881		bigshopping	0.874
innatthecrossroads	caadf	0.885	**talkofweb**	caadf	0.874
	agorams	0.885		agorams	0.874
	bundasgostosas	0.885		bundasgostosas	0.874
	aiesec	0.885		aiesec	0.874
	bigshopping	0.885		bigshopping	0.874
themamamaven	caadf	0.874	**jccnn**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874

Tabela H.4: Tabela Similaridade Base Alexa x 100 Consultas - Parte 4

Tabela Nível de Similaridade Base Alexa x 100 Consultas (Parte 4)					
Consulta	**Script Base Alexa**	**Similaridade**	**Consulta**	**Script Base Alexa**	**Similaridade**
sarahbakesgfree	caadf	0.874	**bonbonbreak**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
mymotherlode	caadf	0.874	**willcookforfriends**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
thecozyapron	caadf	0.874	**beautythroughimperfection**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
iscrapapp	caadf	0.874	**mynewestaddiction**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
forthefamily	caadf	0.874	**littlemissmomma**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
texastypeamom	caadf	0.874	**lifeonvirginiastreet**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
upsda	caadf	0.874	**freewilliamsburg**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
techsmart	caadf	0.874	**natashalh**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
thefirstmess	caadf	0.874	**theysmell**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
happyhealthymama	caadf	0.874	**familyfuninomaha**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
pattern-paradise	caadf	0.874	**thekitchenpaper**	caadf	0.874
	agorams	0.874		agorams	0.874
	bundasgostosas	0.874		bundasgostosas	0.874
	aiesec	0.874		aiesec	0.874
	bigshopping	0.874		bigshopping	0.874
feelingfoodish	caadf	0.874	**naivecookcooks**	bigdatabusiness	0.873
	agorams	0.874		andreastaudt	0.871
	bundasgostosas	0.874		blogdaqualidade	0.871
	aiesec	0.874		camposdojordao	0.871
	bigshopping	0.874		besttemas	0.871
urbangardensweb	caadf	0.874	**quirkychrissy**	brazilkorea	0.872
	agorams	0.874		bhaz	0.864
	bundasgostosas	0.874		cengage	0.864
	aiesec	0.874		bigdatabusiness	0.863
	bigshopping	0.874		andreastaudt	0.861

Tabela H.5: Tabela Similaridade Base Alexa x 100 Consultas - Parte 5

Tabela Nível de Similaridade Base Alexa x 100 Consultas (Parte 5)					
Consulta	**Script Base Alexa**	**Similaridade**	**Consulta**	**Script Base Alexa**	**Similaridade**
catalina-ponor	caadf	0.874	**havadurumux**	annaramalho	0.867
	agorams	0.874		bosontreinamentos	0.848
	bundasgostosas	0.874		bionexo	0.836
	aiesec	0.874		99vidas	0.824
	bigshopping	0.874		animeyes	0.823
yupitsvegan	caadf	0.874	**mobilinkmobile**	cavzodiaco	0.626
	agorams	0.874		casamentos	0.609
	bundasgostosas	0.874		antena1	0.596
	aiesec	0.874		academiadomarketing	0.591
	bigshopping	0.874		cabralsilvaadvogados	0.591
gooddinnermom	caadf	0.874	**najducokoliv**	belezanaweb	0.611
	agorams	0.874		classificados	0.601
	bundasgostosas	0.874		ccine10	0.601
	aiesec	0.874		apdesp	0.601
	bigshopping	0.874		cidadesdomeubrasil	0.597
uglyducklinghouse	caadf	0.874	**neesoku**	bocaonews	0.581
	agorams	0.874		arquitetosdesucesso	0.535
	bundasgostosas	0.874		acontecebotucatu	0.528
	aiesec	0.874		cabanadoleitor	0.522
	bigshopping	0.874		cinejanews	0.512