UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GERAÇÃO DE RÓTULO DE PRIVACIDADE POR PALAVRAS-CHAVES E CASAMENTO DE PADRÕES

DIEGO ROBERTO GONÇALVES DE PONTES

ORIENTADOR: PROF. DR. SERGIO DONIZETTI ZORZO

São Carlos - SP
Junho/2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GERAÇÃO DE RÓTULO DE PRIVACIDADE POR PALAVRAS-CHAVES E CASAMENTO DE PADRÕES

DIEGO ROBERTO GONÇALVES DE PONTES

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Sistemas Distribuídos e Redes de Computadores.
Orientador: Prof. Dr. Sergio Donizetti Zorzo

São Carlos - SP
Junho/2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

P813g

Pontes, Diego Roberto Gonçalves de
Geração de rótulo de privacidade por palavras-chaves e casamento de padrões / Diego Roberto Gonçalves de Pontes. -- São Carlos : UFSCar, 2016.
103 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2016.

1. Política de privacidade. 2. Rótulo de privacidade . 3. Tabela . 4. Casamento de padrões . 5. Privacidade. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de Dissertação de Mestrado do(a) candidato(a) Diego Roberto Gonçalves de Pontes, realizada em 13/07/2016.

Prof. Dr. Sergio Donizetti Zorzo
(UFSCar)

Prof. Dr. Hélio Crestana Guardia
(UFSCar)

Prof. Dr. Marco Paulo Amorim Vieira
(UC / Portugal)

Certifico que a sessão de defesa foi realizada com a participação à distância do membro Prof. Dr. Marco Paulo Amorim Vieira. Depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da comissão examinadora redigido no relatório de defesa do(a) aluno(a) Diego Roberto Gonçalves de Pontes.

Prof. Dr. Sergio Donizetti Zorzo
Coordenador da Comissão Examinadora
(UFSCar)

Dedico este trabalho aos meus pais, minhas maiores inspirações e motivação, sendo sempre eles um porto seguro para os momentos mais difíceis no decorrer do curso de Mestrado.

AGRADECIMENTO

Agradeço primeiramente a Deus, por sempre me manter forte perante as dificuldades da carreira acadêmica. Agradeço aos meus pais, Creusa e Ivan, por sempre me apoiarem e acreditarem em mim. Agradeço ao meu orientador, Sergio Donizetti Zorzo, sempre orientando e mostrando os caminhos certos para a difícil caminhada. Agradeço aos meus amigos de curso Alex Guido, João Moreira, Tiago Rosa, Roan Simões, Anderson Kanegae e principalmente ao André Landi, este sempre me ajudando e sendo como um irmão. Agradeço especialmente a uma pessoa muito querida, Rafaela Sanchioli, que sempre me apoiou, nunca deixou de me incentivar e, além de tudo, sempre acreditou no meu potencial, estando sempre do meu lado nos momentos mais difíceis. Agradeço a todos que participaram direta e indiretamente de mais esta conquista.

"Mais do que máquinas precisamos de humanidade. Mais do que inteligência precisamos de afeição e doçura.
Sem essas virtudes a vida será de violência e tudo estará perdido."

Charles Chaplin

RESUMO

Comumente, os usuários não leem as políticas de privacidade dos serviços online que utilizam. Entre as principais causas estão os textos longos, muitas vezes de difícil compreensão, desestimulando o interesse pela leitura atenciosa e integral. Neste cenário, os usuários, muitas vezes, concordam com os termos sem saber os tipos de dados que estão sendo coletados e o porquê. Esta dissertação discute como o conteúdo das políticas de privacidade pode ser apresentado de forma mais sintética para o usuário, com as informações sobre a coleta e a utilização dos dados sendo exibidas em uma tabela, denominada Rótulo de Privacidade. O Rótulo de Privacidade é uma tabela com linhas nomeadas por termos de coleta de dados e colunas nomeadas por expressões que denotam finalidade das coletas. O conteúdo da tabela informa se a política contempla a coleta de dados para a finalidade especificada. Para ser possível a geração do Rótulo de Privacidade, foi feito um estudo em um conjunto de políticas de privacidade para verificar quais termos mais se repetem nos textos. Para isto foram utilizadas técnicas para encontrar palavras-chave e com estas foram criadas categorias de privacidade. As categorias definem tipos de dados coletados e propósitos da coleta, que no Rótulo de Privacidade são representados pelas células da tabela. Utilizando técnicas de comparação de palavras, uma política de privacidade a ser lida pelo usuário pode ser analisada pela abordagem, extraindo informações importantes por meio das comparações de seus termos com os termos das categorias de privacidade elaboradas. Para cada categoria encontrada na política de privacidade, a informação é ilustrada no Rótulo de Privacidade. Para a avaliação da abordagem proposta, foi desenvolvido um protótipo de uma aplicação, denominada PPMark, que analisa uma particular política de privacidade, extrai as palavras-chave e gera o Rótulo de Privacidade de forma automatizada. As informações extraídas foram analisadas quanto à qualidade utilizando-se três métricas que são empregadas para a avaliação de classificadores, sendo elas precisão, *recall* e *f-measure*. Os resultados mostraram que a abordagem proposta é uma alternativa funcional para o preenchimento do Rótulo de Privacidade e a apresentação das políticas de privacidade. Há evidências de economia de tempo com a leitura e entendimento das políticas, possibilitando suporte para a tomada de decisões.

Palavras-chave: política de privacidade, rótulo de privacidade, tabela, casamento de padrões, palavras-chave, privacidade.

ABSTRACT

Users do not usually read privacy policies from online services. Among the main reasons for that is the fact that such policies are long and commonly hard to understand, which makes the user lose interest in reading them carefully. In this scenario, users are prone to agree to the policies terms without knowing what kind of data is being collected and why. This dissertation discusses how the policies' content may be presented in a more friendly way, showing information about data collection and usage in a table herein called Privacy Label. The Privacy Label is a table with lines named according to data collection terms and columns named according to expressions that reveal how the data is used by the service. The table content shows if the policy collects a particular data to a particular usage. To generate the Privacy Label, a study was made in a set of privacy policies to identify which terms repeat more often along the texts. To do so, we used techniques to find keywords, and from these keywords we were able to create privacy categories. The categories define which kind of data is being collected and why, which are represented by cells in the Privacy Label. Using word comparison techniques, a privacy policy can be analyzed and important information can be extracted by comparing its terms with the terms from the privacy categories. For each category we find, we show it in the Privacy Label. To assess the proposed approach we developed an application prototype, herein called PPMark, that analyzes a particular privacy policy, extract its keywords and generates the Privacy Label automatically. The information extracted was analyzed regarding its quality using three metrics: precision, recall and f-measure. The results show that the approach is a viable functional alternative to generate the Privacy Label and present privacy policies in a friendly manner. There are evidences of time saving by using our approach, which facilitates the process of decision making.

Keywords: privacy policy, privacy label, table, pattern matching, keywords, privacy.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API - *Application Programming Interface*

CPF - Cadastro de Pessoa Física

FN - *False Negative*

FP - *False Positive*

FTC - *Federal Trade Commision*

GT - *Ground Truth*

IHC - Interação Humano Computador

LDA - *Latent Dirichlet Allocation*

LPO - Lógica de Primeira Ordem

MAC - *Media Access Control*

NLP - *Natural Language Processing*

OECD - *Organization for Economic Co-operation and Development*

P3P - *Platform for Privacy Preferences*

PCA - Políticas de Controle de Acesso

PLN - Processamento de Linguagem Natural

RFID - *Radio-Frequency IDentification*

RG - Registro Geral

TF-IDF - *Term Frequency-Inverse Document Frequency*

TP - *True Positive*

SUMÁRIO

Capítulo 1

CONTEXTUALIZAÇÃO

Este capítulo apresenta o contexto no qual o trabalho está inserido e a motivação para o seu desenvolvimento, os objetivos e a forma como está organizado.

1.1 Contextualização

O comércio eletrônico e as redes sociais estão presentes na vida de quase todas as pessoas, o que pode ser facilmente constatado em locais públicos, como shoppings, onde a maioria das pessoas está utilizando dispositivos móveis para acessar serviços on-line ou redes sociais. Mas nem todos os usuários que fazem uso têm a consciência das políticas de privacidade de tais serviços. McDonald e Kranor (2008) afirmam que os usuários, de forma geral, não estão propensos a ler as políticas de privacidade e aceitam os termos sem saber para que suas informações serão utilizadas.

Porém, redes sociais, servidores de e-mails, sites de buscas, serviços de relacionamentos, cada um deles tem uma política de privacidade formulada. Nesses textos são descritos de que forma os usuários fornecem suas informações e como serão utilizadas. Ocorre, todavia, que os usuários não têm o costume de ler toda a política de privacidade do serviço que estão utilizando. Assim, inconscientemente, pode estar consentindo para a utilização de suas informações e elas podem até estar sendo usadas para outros fins, como, por exemplo, o compartilhamento dos dados com outras empresas com o propósito de oferecer propagandas personalizadas.

Acquisti, Brandimarte e Loewensteil (2015) evidenciaram em suas pesquisas que, dentro das culturas sociais, pessoas têm comportamentos diferentes em relação à preocupação com a privacidade e as informações que a tratam como privadas. Além disso, as preocupações com a privacidade podem variar drasticamente para um mesmo indivíduo e para as sociedades ao longo do tempo. A preocupação sobre o que compartilhar e o que manter privado é um dilema universal. Entretanto, a consciência e a preocupação do usuário em relação à sua privacidade possuem influências culturais e regionais. Se esses fatores são preocupações de um indivíduo e da sociedade, deve-se criar uma conscientização a respeito de privacidade. Nas interações entre pessoas e no comércio eletrônico, por exemplo, é necessário saber como os usuários podem ser monitorados, como suas informações estão sendo recolhidas e se elas podem estar sendo compartilhadas com outros indivíduos ou empresas.

Como as políticas de privacidade normalmente são extensas e com linguagem de difícil compreensão para alguns indivíduos, há uma necessidade de facilitar o acesso ao usuário. O trabalho de Kelley et al. (2009) possibilita a apresentação das informações dessas políticas em um formato alternativo, não o textual, mas sim em um formato tabular, tal como os rótulos nutricionais. O formato tabular pode tornar as políticas de privacidade mais fáceis de serem compreendidas, pois são apresentadas de forma mais objetiva, podendo os usuários ficar mais atentos quanto à privacidade de suas informações.

McDonald e Kranor (2008) descreveram o tempo que os norte-americanos gastam lendo políticas de privacidade e fizeram uma previsão de quanto tempo os usuários gastariam para ler todas as políticas de privacidade de todos os serviços de que fazem uso. Os autores também mediram o quanto esse tempo vale em dinheiro e qual seria o prejuízo relacionado, caso os usuários lessem todas as políticas de privacidade. Por outro lado, os autores afirmam que, se houvesse um modelo mais simplificado das políticas de privacidade, os usuários não gastariam tanto tempo para entender o que os termos querem dizer e não os aceitariam de forma inconsciente.

Um dos fatores que influenciam a leitura das políticas de privacidade, segundo Kelley et al. (2010), é a falta de uma padronização de sua escrita. Cada serviço elabora suas políticas de privacidade de forma particular, sem padronização e sem diretrizes na escrita. Alguns serviços tentam ser mais diretos nos termos,

outros já disponibilizam glossários com os jargões utilizados na área de sistemas de informação para auxiliar os usuários.

1.2 Motivação e objetivos

As principais características das dificuldades nas relações entre usuários e políticas de privacidade estão às vezes relacionadas ao tamanho do texto da política, aos termos técnicos que algumas podem apresentar e à falta de tempo para ler todas as políticas de todos os serviços de que fazem uso.

Essas características, como dito anteriormente, fazem com que os usuários aceitem os termos sem mesmo saber quais dados podem ser coletados e para que são utilizados. A tabela elaborada por Kelley et al. (2009) para a apresentação das políticas de privacidade trouxe resultados sobre os usuários, melhorando suas compreensões sobre tais termos.

O mecanismo proposto pelos autores para a geração das tabelas utiliza a plataforma *Platform for Privacy Preferences*[1] (P3P), que não é utilizada pela maioria dos serviços, sendo que, na maioria dos casos, as políticas de privacidade são apresentadas em linguagem natural (formato textual).

Sendo assim, a motivação desta pesquisa foi desenvolver uma solução para extrair informações sobre coleta e utilização dos dados dos usuários e apresentar tais informações em formato de tabela, denominada neste trabalho como Rótulo de Privacidade.

Para é apresentar as políticas de privacidade em formato de tabela, construiu-se um catálogo de palavras-chaves de privacidade sobre um conjunto de textos e com um formalismo para casamento de padrões, para que, seja possível analisar uma política e gerar o Rótulo de Privacidade.

De modo geral, os objetivos do trabalho são: gerar um catálogo com os termos de privacidade utilizados nas políticas textuais; criar categorias de privacidade para tais palavras; definir um formalismo para ocorrências de casamento de padrões; implementar um protótipo de uma aplicação para analisar políticas de

[1] https://www.w3.org/P3P/

privacidade textual e, utilizando o protótipo, avaliar a extração de informações sobre coleta e utilização de dados descritos nas políticas.

1.3 Metodologia de desenvolvimento do trabalho

Como as políticas de privacidade são apresentadas em linguagem natural pelos serviços on-line e os usuários podem ter dificuldades para o seu entendimento, aceitam os termos sem estar conscientes de quais dados estão sendo coletados e para quais propósitos podem ser utilizados.

Após identificar o problema de apresentação da política de privacidade para os usuários, citado anteriormente, foram realizadas pesquisas tendo como objetivo encontrar técnicas e soluções para facilitar a apresentação das políticas de privacidade. As pesquisas foram realizadas em máquinas de indexação como IEEE, ACM, Scopus, ACM e o Google Acadêmico. As palavras-chaves para filtrar os artigos das máquinas de indexação foram: *privacy policy, extracting information, natural language processing (nlp), patter matching*. A busca retornou vários trabalhos, porém foram selecionados apenas os que contribuíram diretamente com o problema.

Para tornar possível a avaliação da proposta deste trabalho foi implementado um sistema para gerar Rótulos de Privacidade, chamado PPMark. Este sistema foi implementando para observar quais características de coleta e utilização de dados podem ser extraídas dos textos de forma automática. Para mensurar essa avaliação, foi calculada a precisão com que a técnica consegue recuperar as informações de interesse (PEREZ-CASTILLO et al., 2011; SEBASTIANI, 2002).

O teste de precisão da extração de informações foi executado sobre um conjunto de políticas de privacidade. A metodologia para avaliação da precisão se deu da seguinte maneira: (i) seleção do corpus de políticas de privacidade para realização da avaliação; (ii) análise e geração manual dos rótulos feita pelo especialista; (iii) análise e geração do rótulo de privacidade feita pela PPMark; e (iv) comparação dos rótulos obtidos pelo especialista e PPMark.

No desenvolvimento do protótipo utilizou-se a linguagem de programação Java. A escolha da linguagem Java se deu pela familiaridade do autor com a

linguagem e as várias implementações das técnicas utilizadas neste trabalho disponíveis para utilização. A ferramenta pode, porém, ser desenvolvida em qualquer outra linguagem, bastando apenas implementar o formalismo proposto.

1.4 Organização do trabalho

O presente trabalho está organizado e apresentado da seguinte forma: no Capítulo 2, vê-se os conceitos de privacidade e técnicas utilizadas no seu desenvolvimento. No Capítulo 3, estão tratados os trabalhos relacionados, suas características e contribuições. No Capítulo 4, trata-se de como o Rótulo de Privacidade pode ser gerado. Capítulo 5, apresenta o protótipo da aplicação PPMark, que é capaz de gerar os Rótulos de Privacidade. No Capítulo 6, a avaliação da técnica e os resultados obtidos são apresentados. No Capítulo 7, estão as conclusões e, por fim, as referências bibliográficas utilizadas.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Neste capitulo são apresentados conceitos sobre privacidade, coletas de dados, compartilhamento de dados com terceiros, relação entre usuários e políticas de privacidade, como são elaboradas as políticas e por fim técnicas de extração de informações em documentos.

2.1 Considerações iniciais

Os serviços on-line que utilizam informações dos usuários para prestar seus serviços devem conter políticas de privacidade disponíveis para acesso. Estas políticas de privacidade contêm termos sobre a coleta, utilização e armazenamento das informações. Porém, segundo McDonald e Cranor (2008), os usuários podem ter dificuldades de entendimento desses termos citados nas políticas ou não dispor de tempo para a sua leitura.

Diante disso, neste capítulo são abordados conceitos sobre a privacidade no contexto digital, políticas de privacidade, os conflitos entre usuários e políticas de privacidade e técnicas para localizar palavras-chaves sobre privacidade nos textos das políticas.

2.2 Privacidade no contexto digital

Para o desenvolvimento deste trabalho, a privacidade pode ser definida como o direito que o indivíduo tem de compartilhar suas informações pessoais seletivamente com outros indivíduos (SMITH e XU, 2011). Westin (1968) abordou a privacidade como sendo o direito do indivíduo de determinar quais informações sobre si mesmo podem ser compartilhadas com outros, como os dados podem ser coletados e em que poderão ser usados. Essas definições abordam a privacidade de

forma genérica. No entanto, é possível que cada pessoa tenha sua própria concepção sobre o que é privacidade. Alguns indivíduos não ficam incomodados em compartilhar sua localização. Já outros já não gostam de ter essa informação divulgada.

Um exemplo disso pode ser dado por dois usuários distintos. Um usuário sempre viaja de avião, então busca constantemente melhores preços. Quando acessa um serviço de venda de passagens, os seus dados de navegação são coletados e ele receberá promoções de passagens em seu e-mail. Ou seja: para ele a propaganda de passagens não é invasiva, pois ele pode receber uma promoção com o preço mais satisfatório da passagem, sendo, assim, um fator benéfico no compartilhamento de dados de navegação. Porém, se um usuário, apenas por curiosidade, fizer uma consulta de valores de passagens de avião, e se seus dados de navegação são coletados, ele começará a receber promoções sobre passagens, sendo que o desejo inicial do usuário era apenas ver quanto custaria uma passagem.

Pearson (2009) trata a privacidade identificando os dados de identidade pessoal e os dados sensíveis, bem como os dispositivos exclusivos utilizados e o uso que será feito dos dados coletados. O autor descreve quatro tipos de identificação: (i) informações de identificação pessoal: qualquer informação que possa ser utilizada para identificar ou localizar um indivíduo, tal como nome, endereço ou ainda informações que podem relacionar um dado com um indivíduo, como, por exemplo, número de cartão de crédito, código postal ou até mesmo endereço de protocolo de internet (PI); (ii) informações sensíveis: são informações sobre religião, raça, saúde, orientação sexual, filiação sindical ou quaisquer outras informações que são consideradas privadas; (iii) uso de dados: são informações que podem ser coletadas de dispositivos pessoais, como histórico de navegação do browser do usuário, por meio do qual se podem traçar hábitos de visualizações de conteúdos digitais; (iv) identidade de dispositivos exclusivos: são informações que podem ser coletadas exclusivamente baseadas em um dispositivo pessoal, como, por exemplo, endereço de protocolo de internet, identificação de rádio frequência (RFID) e identidades de hardware único, como endereço *Media Access Control* (MAC).

Com outro entendimento, Ghani e Sidek (2008) afirmam que algumas informações pessoais são sensíveis, mas não necessariamente precisam ser

pessoais. Os autores classificam quatro tipos de dados: (i) dados pessoais: nome, endereço, número de telefone; (ii) dados sensíveis: origem racial, étnica, religiosa, filiação em partidos políticos, dados de histórico de saúde; (iii) dados de identificação: DNA, número de documento pessoal (identidade); (iv) dados anônimos: todos os dados que não podem ser vinculados diretamente com um indivíduo, como sexo ou tipo de doença.

Já no texto do projeto de lei de proteção de dados pessoais[2], dados pessoais e dados sensíveis foram definidos como: 1. dados pessoais: dados relacionados à pessoa natural identificada ou identificável, inclusive a partir de números identificativos, dados locacionais ou identificadores eletrônicos; e 2. dados sensíveis: dados pessoais que revelem a origem racial ou étnica, as convicções religiosas, filosóficas ou morais, as opiniões políticas, a filiação a sindicatos ou organizações de caráter religioso, filosófico ou político, dados referentes à saúde ou à vida sexual, bem como dados genéticos.

Tanto os dados pessoais como os sensíveis são fundamentais para a identificação de uma pessoa. Quando as lojas de comércio eletrônico vão disponibilizar seus serviços, o usuário necessita fazer um cadastro, no qual são solicitadas informações pessoais, tais como nome completo, cadastro de pessoa física (CPF), registro geral (RG) e endereço para fins tributários e, dependendo do serviço, utiliza-se o endereço para a entrega de um produto. Tais dados são valiosos tanto para os usuários quanto para as empresas que prestam serviços on-line. Dessa forma, os serviços que necessitam de cadastro dos usuários e aqueles em que não há necessidade do usuário fornecer informações explicitamente, utilizam-se de técnicas diferentes para adquirir informações.

Esses dados são importantes para o varejo, pois com técnicas de mineração de dados, as empresas podem facilmente extrair dados pessoais dos consumidores a partir apenas dos seus nomes e endereços de correio eletrônico. Os dados podem conter históricos de compras on-line e/ou off-line, preferências de músicas, filmes, roupas e até programas de fidelização dos quais o consumidor faz parte. Com posse dessas informações, as empresas podem oferecer propaganda personalizada para cada consumidor, baseando-se em seu perfil de compra, navegação e preferências pessoais.

[2]https://participacao.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/

Os consumidores devem ficar atentos aos textos das políticas de privacidade dos serviços on-line que utilizam. A forma como os dados são coletados e armazenados devem estar descritos nas políticas de privacidade. Além da coleta das informações dos usuários e armazenamento, algumas empresas compartilham os dados com terceiros, seja por parceria publicitária ou mesmo por venda de informações.

Conger, Pratt e Loch (2013) fizeram um levantamento sobre a privacidade da informação pessoal e descreveram o compartilhamento em quatro tipos de entidades sobre os dados dos usuários, conforme mostrado na Figura 2-1.

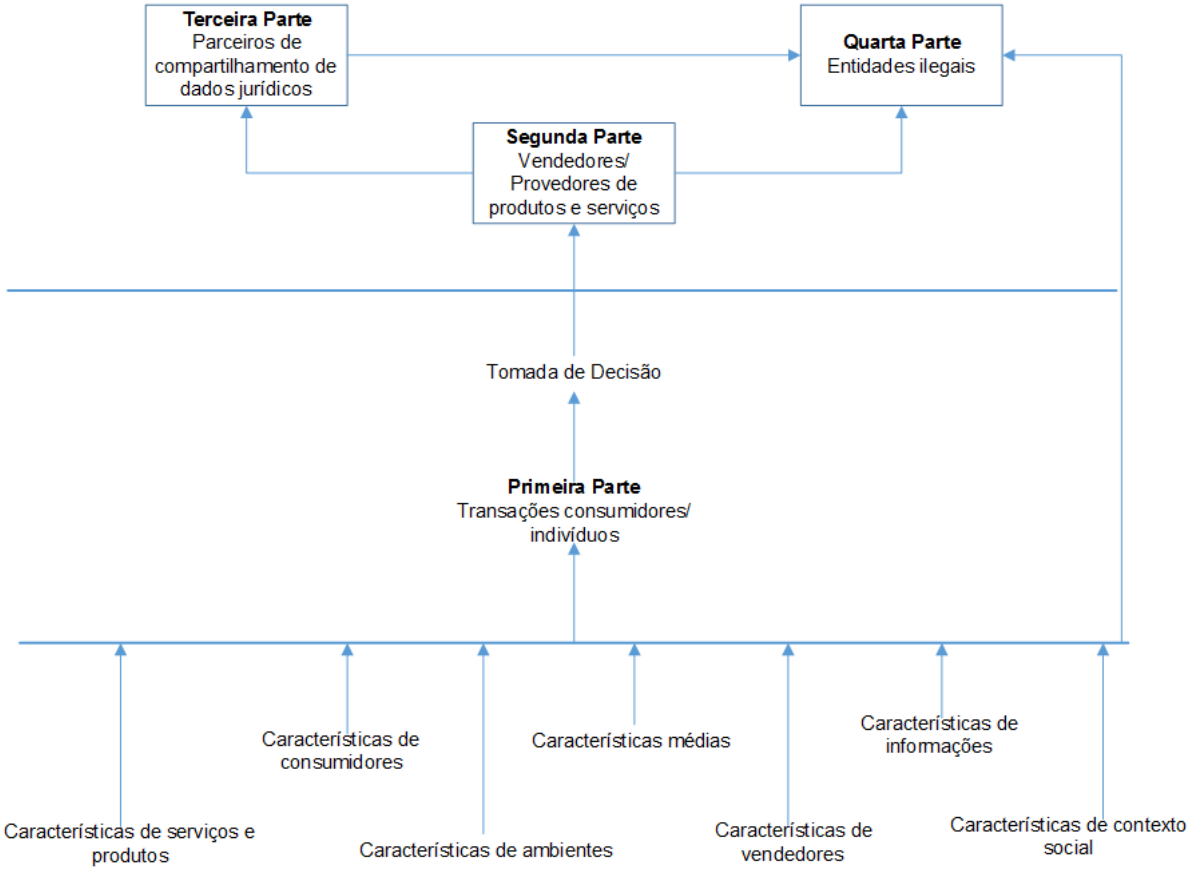

Figura 2-1: Modelo de privacidade expandido adaptado de Conger, Pratt e Loch (2013)

Na Figura 2-1 é possível identificar quatro interessados nos dados. Como primeira parte tem-se o consumidor ou indivíduo que tem a propriedade sobre o dado e forneceu as informações para poder usar o serviço on-line. A segunda parte é aquela que vende ou fornece serviços, que necessita de informações pessoais para realizar alguma operação, desde compras em sites de vendas até

transferências bancárias. Na transição dos dados da primeira para a segunda parte, o usuário dá permissão para a segunda parte fazer uso de suas informações, confiando-lhe, assim, seus dados.

Os usuários podem não estar cientes de que uma vez feita uma troca de interesses para a segunda parte, os registros da transação podem automaticamente ser compartilhados com parceiros, sendo estes chamados de terceira parte. Neste momento, o usuário não tem mais controle sobre seus dados, ficando à disposição de terceiros a utilização das informações, o que pode ocorrer sem o seu conhecimento. A terceira parte engloba várias entidades: uma empresa terceirizada que consulta informações de crédito de um indivíduo, entidades regulamentadas pelo governo ou até mesmo empresas que fazem propaganda direcionada e personalizada com base nos dados compartilhados pela segunda parte. A terceira parte pode ser conhecida, mas não entra nas decisões de compartilhamento de informações dos usuários no ato da transação com algum serviço, ou seja, a terceira parte está oculta entre o usuário e o serviço que está utilizando. Segundo Conger, Pratt e Loch (2013), a quarta parte é a que oferece mais riscos aos dados dos usuários, podendo ser hackers ilegais, ladrões ou funcionários de terceiros que violem a política da empresa.

A ameaça de violação de privacidade do usuário está fortemente presente nas transações da primeira parte, descrita na Figura 2-1, justificando-se que muitas descrições da política de privacidade não são entendidas ou observadas pelos usuários.

2.3 Políticas de privacidade

Políticas de privacidade são documentos (contratos) que descrevem termos para garantir a privacidade das informações dos usuários. Internacionalmente existem duas organizações com o objetivo de visar a proteção de privacidade dos usuários em serviços on-line: *Organization for Economic Co-operationand Development*[3] (OECD) e *Federal Trade Commision*[4] (FTC).

[3] http://www.oecd.org/
[4]https://www.ftc.gov/

A OECD é direcionada para a proteção da privacidade dos usuários, contendo informações específicas para sua privacidade. Ela estabelece de que maneira as informações pessoais dos usuários devem ser protegidas. Os principais embasamentos da OECD são: (i) conter um limite de coleta de dados; (ii) os dados pessoais dos usuários devem ser autênticos para os objetivos para os quais serão utilizados, ou seja, os usuários devem fornecer dados verdadeiros; (iii) especificar o objetivo da coleta dos dados antes mesmo de efetuar a coleta de informações; (iv) limitar o uso dos dados coletados, como, por exemplo, fornecer dados para terceiros. Se isso for ocorrer, deve ser explicado o motivo do compartilhamento dos dados com terceiros no texto da política de privacidade.

Já a FTC é direcionada a apoiar os cuidados da privacidade e da vida econômica dos cidadãos, criando leis em favor da segurança dos dados pessoais, aplicando investigações contra criminosos, cujo foco é coibir fraudes financeiras, e fornecendo orientações para os usuários tomarem decisões de compras. Segundo Pitofsky et al. (2000), a FTC define princípios para práticas justas de privacidade, baseando-se em uma legislação que protege informações pessoais quando coletadas. Os embasamentos da FTC seguem os mesmos da OECD e mais algumas contribuições: (i) os sites devem atualizar os usuários sobre a coleta de seus dados; (ii) negociar a privacidade na coleta com o site, ou seja, poder escolher o que pode ou não ser utilizado; (iii) fornecer acesso aos dados que foram coletados, dando o direito de poder atualizar ou corrigir algumas informações; (iv) os serviços devem ser responsáveis por proteger com segurança todas as informações coletadas dos usuários.

Zorzo e Lobato (2007) conduziram um estudo de caso no qual o foco foi aplicar uma avaliação por inspeção em sites brasileiros de e-commerce para verificar as características relacionadas à privacidade. A partir desse estudo, outro trabalho foi conduzido por Lobato e Zorzo (2007). Os autores constataram que apenas alguns sites têm uma política de privacidade que utiliza regras e deveres claros aos usuários. No entanto, essas políticas não possuem um padrão na escrita dos textos. O estudo de caso buscou elencar as características de privacidade e personalização. Lobato e Zorzo propuseram uma padronização para políticas de privacidade utilizadas pelos sites, analisando vários modelos de padrões. Relatam que existem algumas técnicas que podem ser utilizadas de modo a tornar os usuários mais conscientes sobre as políticas, que são: (i) devem ser exibidas aos

usuários de maneira que sejam claramente compreendidas; (ii) os usuários devem ter a escolha de negociação para o uso de seus dados; (iii) a cada alteração em uma política de privacidade, os usuários devem ser capazes de visualizá-las. Usando coleções de padrões definidos e fundamentos das instituições OECD e FTC, os autores estruturaram uma base para o desenvolvimento de uma política de privacidade para ser tomada como exemplo.

De acordo com Lobato e Zorzo (2007a), o modelo apresentado pode ser utilizado para trazer facilidades aos serviços on-line na definição de suas políticas de privacidade. Afirmam que tal proposta traz benefícios aos usuários, pois o conteúdo das políticas de privacidade é definido de maneira mais clara e objetiva e em uma linguagem que o usuário entende, aumentando a confiabilidade entre ele e serviços on-line, mas que dificilmente, em termos comerciais, as empresas estariam dispostas a adotar tais padrões.

Em esfera nacional, o governo federal aprovou o Marco Civil da Internet[5] com o propósito de regulamentar a neutralidade da rede de acesso, a guarda de registros de conexões e a retirada de conteúdos e responsabilidades. O projeto começou a ser elaborado em 2009 e sancionado em 2014. A lei dita regras tanto para usuários quanto para fornecedores de serviços. O ponto forte do Marco Civil, segundo reportagem publicada[6]no site da Globo, é a garantia de neutralidade de rede. Esta garantia diz respeito ao fato de que os provedores de internet ficam proibidos de ofertar conexões diferenciadas a partir do conteúdo que o usuário for acessar. Um exemplo: provedores já forneceram conexão para e-mails na qual o acesso poderia ser liberado somente na leitura e envio de mensagens por correio eletrônico. Os provedores já não podem mais fornecer esse tipo de conexão. Se o usuário tem acesso à internet pelo provedor, então ele pode acessar qualquer conteúdo disponível na web. Entretanto, o Marco Civil tem um ponto que pode ser discutido por um longo tempo.

A lei diz que todos os registros de navegação dos usuários devem ser mantidos por um ano e estar disponíveis para o governo a qualquer momento, e os provedores têm de manter tais registros em total sigilo e proteção. Porém, os registros estão relacionados somente ao endereço de IP, data e horas inicial e final

[5] Texto completo disponível em: http://www.camara.gov.br/sileg/integras/912989.pdf

[6] http://g1.globo.com/tecnologia/noticia/2014/06/marco-civil-da-internet-entra-em-vigor-nesta-segunda-feira-23.html

da conexão. Esses registros são armazenados, segundo a lei, para combater crimes virtuais, como pedofilia e outras formas de crime cibernético. O documento ainda fixa princípios de privacidade sobre os dados que os usuários fornecem aos provedores, para que não possam ser usados diferentemente do que estabelece a política de privacidade do serviço. Já a liberdade de expressão e retirada de conteúdo do ar é um ponto favorável para os provedores. O Marco Civil também pretende defender a liberdade de expressão no mundo virtual, determinando que um conteúdo só pode ser retirado do ar após uma ordem judicial e que os provedores não poderão ser responsabilizados por conteúdos ofensivos postados em seus serviços pelos usuários.

Após o Marco Civil entrar em vigor, houve mais discussões sobre a privacidade dos dados dos usuários. Essas discussões chegaram à Câmara dos Deputados, que começou a elaborar um novo Projeto de Lei de Proteção de Dados[7]. O novo Projeto de Lei está sendo elaborado e usuários podem dar sugestões para o texto até a sua aprovação. O Projeto tem como foco estabelecer diretrizes para as empresas em relação aos dados fornecidos ou coletados dos usuários e está diretamente relacionado aos deveres que os serviços on-line publicam em seus textos de políticas de privacidade.

O ato de coletar, utilizar, acessar, reproduzir compartilhamento é denominado no texto do Projeto de Lei como tratamento de dados. Sobre o tratamento de dados é explicito que, para ocorrer, deve haver o consentimento do usuário, e que o tratamento deve ter uma finalidade determinada. O texto também informa que o tratamento de dados deve ser destacado das demais cláusulas contratuais. Se houver tratamento de dados, deve ser mostrado na política de privacidade o tempo que o dado será armazenado, utilizado e se será compartilhado. É vedado qualquer tipo de compartilhamento de dados pelo serviço on-line, a não ser que haja consentimento explicito do usuário.

Com o Marco Civil da Internet e o novo Projeto de Lei de Proteção de Dados, o Brasil fortalece suas leis para ambientes virtuais. Em princípio a preocupação era voltada mais para as leis contra cibercrimes. Entretanto, com a grande venda de dados a terceiros por empresas, focou-se em definir leis que fazem proteção dos dados dos usuários. Mas, mesmo com o Marco Civil e o novo projeto de lei, as

[7]http://pensando.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/

definições de coleta, uso, compartilhamento e divulgação estão descritas nas políticas de privacidade. Agora cabe ao usuário, antes de fornecer dados para um determinado serviço on-line, fazer uma investigação quanto às suas políticas de privacidade.

2.4 Conflitos entre usuários e as políticas de privacidade

Smit, Noort e Voorveld (2014) conduziram um experimento que consistiu em uma entrevista com 2022 usuários. A entrevista foi relacionada à publicidade baseada no comportamento on-line, explicando se as informações coletadas para a publicidade eram direcionadas somente para o serviço que estavam utilizando, se eram compartilhadas com terceiros, entre outros aspectos. Dada a explicação sobre a utilização das informações, dentre os entrevistados, apenas 7% estariam dispostos a compartilhar informações com terceiros. Com a entrevista conduzida pelos autores, na qual foram elucidados conceitos sobre a utilização das informações, pode-se notar que é necessária uma explicação clara e direta sobre as políticas de privacidade dos serviços on-line, de tal forma que, lida a política pelo usuário e se ela estiver clara, coerente e direta, a sua tomada de decisão possa ser influenciada.

Diante de coleta, compartilhamento e armazenamento dos dados dos consumidores, as políticas de privacidade têm a finalidade de informar aos usuários como seus dados serão cuidados quando divulgados para eles terem acesso ao serviço. Porém, os textos de políticas de privacidade são extensos, com jargões técnicos, sem padrões de escrita e muito complexos. Para o usuário usufruir dos serviços on-line, ele deve ler os termos de política de privacidade e decidir se vai aceitar os termos ou não. Caso o usuário não aceite os termos, ele fica impossibilitado, muitas vezes de usufruir das funcionalidades prestadas pelo serviço. Atualmente, nas implementações dos serviços on-line, o usuário é incapaz de negociar sua privacidade, ou seja, é incapaz de escolher quais dados ele permite que sejam coletados, como devem ser usados. Ou ele fornece os dados que são necessários e usufrui dos serviços ou não fornece e não tem acesso aos serviços.

Jensen e Potts (2004) afirmam que a falta da opção de negociação da privacidade para os usuários é um ponto falho nos serviços de comércio eletrônico.

Acquisti, Brandimarte e Loewenstein (2015) afirmam que usuários comuns não leem as políticas de privacidade e logo aceitam os termos sem saber o que será feito com seus dados após a divulgação para o serviço. Os autores afirmam que um desafio muito importante da área de Interação Humano Computador (IHC) é como transmitir uma série de informações complicadas e não sobrecarregar os usuários.

Jensen e Potts (2004) e Reidenberg et al. (2015) afirmam que uma política pode ser clara, dependendo do entendimento do público-alvo e das habilidades de leitura e compreensão. Porém, o público pode ter certa habilidade de leitura e compreensão, mas em relação a sistema de informação seu conhecimento pode não ser tão satisfatório para o entendimento dos termos das políticas de privacidade.

Como as políticas de privacidade não possuem uma regulamentação definida e não têm padrões explícitos de conteúdo e escrita, é necessário que os usuários leiam toda a política de privacidade de todos os serviços que forem usar, isto caso tenham interesse. No entanto, isso normalmente não acontece, fazendo com que o consentimento do acesso ao serviço se torne implícito.

Nos textos das políticas de privacidade existem vários problemas de estrutura e conteúdo (LOBATO e ZORZO, 2007a). O contexto legal das políticas as torna inutilizáveis como apoio às decisões para um usuário preocupado com a privacidade (MCDONALD e CRANOR, 2008). É colocada uma imensa carga no indivíduo final, pois não são feitas notificações adequadas quanto a alterações nos textos ou à apresentação das políticas de privacidade em uma linguagem que qualquer usuário possa ler, compreender o conteúdo, o que está sendo requisitado dele e de que forma serão utilizadas suas informações.

As políticas de privacidade têm um fator essencial, pois elas notificam os usuários se seus dados estão sendo coletados, armazenados e/ou divulgados. Os serviços on-line, em suas políticas de privacidade, de forma clara ou não, descrevem esses avisos em seus textos de políticas de privacidade. O indivíduo que tem consciência sobre privacidade deve ler todo o texto de política de privacidade para saber o que será feito com suas informações após a divulgação para a utilização dos serviços.

Segundo McDonald e Cranor (2008), o tempo para ler as políticas de privacidade é uma forma de pagamento, ou seja, ao invés de receber pagamentos

para revelar suas informações, os usuários devem pagar com seu tempo lendo e investigando as políticas para poder manter sua privacidade. Os autores estimam que a leitura de políticas de privacidade custa ao usuário aproximadamente 201 horas por ano. Em valores comerciais isto custa 3.534 dólares americanos por ano para cada usuário de internet nos EUA. Em nível nacional, se os americanos lerem todas as políticas de privacidade dos serviços on-line de que fazem uso, palavra por palavra, o valor estimado do tempo perdido é de 781 bilhões de dólares anualmente.

As empresas sempre consideram que seus usuários devem ler as políticas de privacidade, e se eles não o fazem, é um sinal de falta de preocupação com a privacidade. Por outro lado, usuários alegam que os sites precisam fazer um melhor trabalho para poder transmitir suas práticas de forma mais útil, sutil, o que inclui diminuir o tempo de leitura das políticas de privacidade. Com uma atenção da área de IHC sobre as políticas de privacidade, como mencionado no trabalho de Acquisti, Brandimarte e Loewenstein (2015), seria fundamental uma padronização nos textos de políticas de privacidade e uma forma mais sofisticada e facilitada para apresenta-las aos usuários. Com isso os usuários perderiam menos tempo na leitura, teriam consciência de sua privacidade naquele serviço on-line e seriam auxiliados nas tomadas de decisões relacionadas à privacidade.

2.5 Palavras frequentes nos textos de políticas de privacidade

Conforme já foi abordado, as políticas de privacidade não seguem um padrão definido de escrita, ficando a critério do serviço on-line elaborar e divulgar suas políticas. Porém, apesar de não terem um padrão definido, as políticas de privacidade estão relacionadas à coleta e utilização de informações dos usuários, além de terem um contexto único e definido: a privacidade. Sendo assim, pode-se afirmar que as políticas possuem termos/palavras comuns entre elas. Para reconhecer esses termos/palavras foi feito um levantamento nas máquinas de busca para encontrar as principais técnicas utilizadas em mineração de texto, capazes de agrupar termos/palavras em um conjunto de textos.

No levantamento feito sobre as técnicas, foram encontradas três principais. Como o enfoque deste trabalho não está nas técnicas de aprendizagem de máquina,

inteligência artificial ou mineração de texto, as três técnicas serão abordadas brevemente nas subseções seguintes. São elas: Frequência do Termo-Inverso da Frequência nos Documentos, *Latent Dirichlet Allocation* (LDA) e por fim o Modelo Skip-gram.

2.5.1 Frequência do Termo-Inverso da Frequência nos Documentos

A técnica Frequência do Termo-Inverso da Frequência nos Documentos, do inglês *Term Frequency-Inverse Document Frequency* (TF-IDF), é utilizada para encontrar os termos mais frequentes em um determinado corpus[8]. Sua aplicabilidade é direcionada a certo conjunto de documentos sobre um tema específico, que neste trabalho será uma coleção de textos de políticas de privacidade.

Essa técnica é utilizada para encontrar palavras-chaves em um grande conjunto de textos. Em um texto pode-se certificar que palavras como "a", "o", "um", "uma" são palavras normalmente repetidas ao longo do texto. A técnica TF-IDF visa não contabilizar tais palavras e letras e sim apenas as palavras-chaves que são relevantes para o tema em consulta. Para identificar palavras-chaves, é necessário verificar não somente quantas vezes uma determinada palavra ocorre em um determinado documento, mas também com que frequência cada palavra ocorre em outros textos. A técnica é um algoritmo semiempírico cuja primeira proposta foi dada por Jones (SPÄRCK JONES, 1972) e vários estudos foram desenvolvidos sobre a ideia original. Ramos, Eden e Edu (2003) descreveram o cálculo do TF-IDF por

$$w_d = f_{w,d} * \log\left(\frac{|D|}{f_{w,D}}\right) \quad (1)$$

onde $f_{w,d}$ é igual ao número de vezes em que w aparece em um documento; d,D é o tamanho do corpus e $f_{w,D}$ é o número de documentos do corpus D, que contém o termo w. As palavras com os maiores valores são selecionadas como palavras mais frequentes e de relevância. A técnica aplica a Equação 1 sobre um corpus, uma vez dados uma palavra, conjuntos de palavras ou frases chaves sobre o tema.

[8] Coletânea ou conjunto de documentos sobre determinado tema.

2.5.2 *Latent Dirichlet Allocation*

A técnica *Latent Dirichlet Allocation* (LDA) foi desenvolvida por Blei, NG e Jordan (2003). A técnica aplica cálculos probabilísticos para encontrar o modelo de temas predefinidos em novos documentos a serem analisados. A equação para encontrar os temas nos documentos é mostrada na equação a seguir.

$$P(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{k=1}^{K} P(\beta_k) \prod_{d=1}^{D} P(\theta_d) (\prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n}) \mid \beta_{1:K}, Z_{d,n})$$

Na equação, $\beta_{1:k}$ representa todos os tópicos em todos os documentos e $W_{d,n}$ representa uma única palavra observada em um documento específico. O θ_d representa as proporções de tópicos por documento d. Com isso, para uma palavra que está sendo verificada, $W_{d,n}$ depende do conjunto de todos os tópicos $\beta_{1:k}$ e a distribuição dos tópicos é dado por um documento $Z_{d,n}$.

Para a execução da LDA são necessárias duas entradas, sendo um conjunto de documentos (corpus) e um número de tópicos, representado na equação por K, que tem a função de controlar a granularidade da modelagem de tópicos. LDA tem em sua equação duas incógnitas principais: a proporção de tópicos por documento θ e a probabilidade de palavras de tópicos β. Segundo o critério de probabilidade máxima, seleciona-se tanto θ e β com valores máximo de probabilidade para os dados observados $P(W|\theta, \beta)$. Aplicando essa técnica, o processo consegue achar palavras do texto que se encontram dentro do conjunto de tópicos.

2.5.3 Modelo Skip-gram

Mikolov et al. (2013) introduziram o modelo Skip-gram para aumentar a qualidade de aprendizagem para representação de palavras em vetores. O modelo utiliza redes neurais para fazer um treinamento prévio em um grande conjunto de palavras. O objetivo do treinamento é encontrar representações de palavras que

podem ser úteis para prever palavras vizinhas em uma frase ou em um documento. Formalmente, é dada uma sequência de palavras $w_1, w_2, w_3, \ldots, w_T$, objetivando maximizar a média do log:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

onde c é o tamanho do contexto do treinamento. A equação acima seleciona uma palavra central e calcula seus vizinhos. Dessa forma é possível, após o treinamento, dada uma palavra, prever quais são os vizinhos mais próximos (MIKOLOV et al., 2013). A Figura 2-2 ilustra o funcionamento do modelo skip-gram.

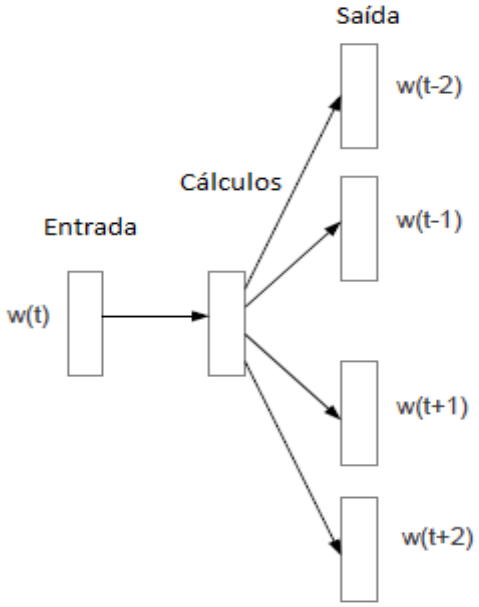

Figura 2-2: Funcionamento do modelo Skip-gram adaptado de Mikolov et al. (2013).

De acordo com a Figura 2-2, dada uma entrada, são executados os cálculos de previsão e como saída têm-se os termos mais próximos da entrada.

2.6 Localizar padrões em textos de políticas de privacidade

Conhecendo os termos/palavras mais comuns nas políticas de privacidade, como descobrir se alguns ou todos desses termos/palavras estão em uma nova política?

Para localizar palavras já conhecidas em um novo texto, existem técnicas que fazem comparações desses termos/palavras com as palavras do novo texto. Essas técnicas são utilizadas por editores de textos (localizar), máquinas de buscas, padrões de DNA e vários outros segmentos. Essas técnicas são usualmente chamadas técnicas de "Casamento de padrões" ou, no inglês, "*Pattern Matching*".

Susik, Grabowski e Fredriksson (2014) definem formalmente casamento de padrão como: dado um texto T de tamanho n e um conjunto de padrões $P = \{P1, \dots, Pr\}$, onde o alfabeto é de tamanho σ, encontrar ocorrências dos padrões em T. Técnicas de casamento de padrões são métodos para verificar se há presença de algum padrão, dada uma base de conhecimento, em um conjunto de dados. Segundo os autores, as técnicas mais utilizadas para casamento de padrões são os algoritmos Boyer-Moore e Rabin-Karp. A seguir serão descritos breves conceitos dessas técnicas.

2.6.1 Algoritmo Boyer-Moore

O algoritmo Boyer-Moore foi desenvolvido pela dupla Robert S. Boyer e J. Strother Moore (1977). Neste algoritmo, são feitas comparações dos caracteres de um padrão com os caracteres de uma *string*. A comparação do padrão com a *string* é feita da direita para esquerda, sendo que, se ocorrer alguma falta de combinação, o algoritmo faz um deslocamento baseado no caractere que gerou essa falta. O exemplo ilustrado na Figura 2-3 apresenta o funcionamento do deslocamento feito pelo algoritmo e a ocorrência do casamento de padrão.

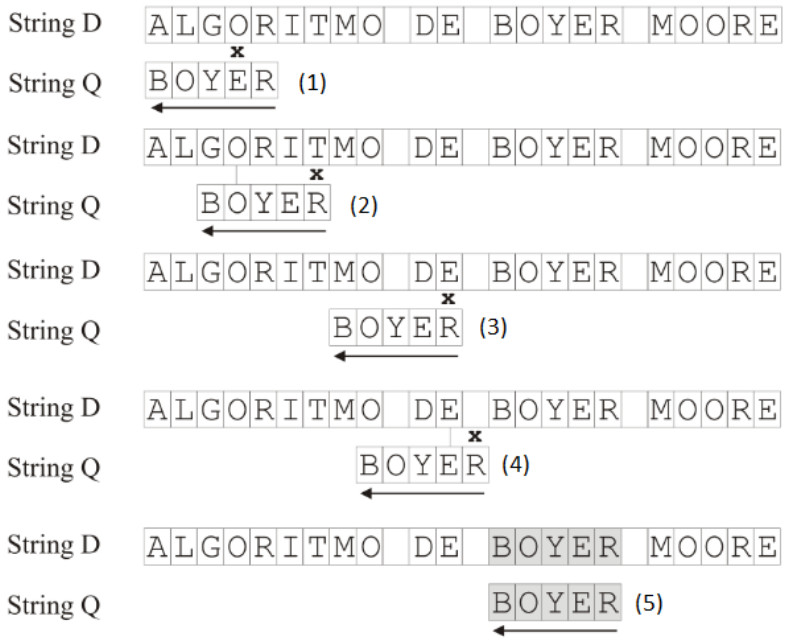

Figura 2-3: Exemplo do funcionamento do deslocamento do algoritmo e a ocorrência do casamento de padrão (MEDEIROS VANDERLEI, 2006)

De acordo com a Figura 2-3, são apresentadas duas *strings*, sendo Q e D, onde D representa a *string* para verificar se há algum padrão e D representa o padrão. No exemplo ilustrado foram necessárias cinco interações para localizar o padrão. O funcionamento é da seguinte maneira: inicialmente D é alinhado com Q. O algoritmo começa a comparar os caracteres da direita para a esquerda, conforme a primeira interação. A comparação do primeiro caractere ocorre com sucesso (‘R’=’R’), porém ocorre uma falta no próximo caractere (‘O’≠’E’). Entretanto, o caractere no qual ocorreu a falta está presente em Q. Desta forma, o algoritmo alinha os caracteres, aquele no qual ocorreu a falta com o que está no padrão, deslocando duas posições, conforme demonstrado na segunda interação. É comparado o primeiro caractere da direita para a esquerda, sendo (‘T’≠’R’). Ocorreu uma falta e o caractere onde ela ocorreu não está presente em Q, então se pode deslocar o padrão todo em 5 posições. Na terceira interação, na primeira comparação ocorre uma falta (‘E’≠’R’), porém ‘E’ está presente em Q, assim o algoritmo alinha o caractere em que ocorreu a falta com o caractere de Q. A lógica continua igualmente para todos os caracteres de D, finalizando com a localização do padrão ou com o fim da cadeia D.

Há diversos trabalhos sobre casamento de padrões. Algumas técnicas fazem busca caractere por caractere, como descrito nos trabalhos de Cantone e Faro (2005), Nebel (2006), Sustik e Moore (2007), Kalsi, Peltola e Tarhio (2008), enquanto outras técnicas já fazem utilização de expressões regulares e autômatos, como proposto por Watson (1996).

É um algoritmo de força bruta, que faz comparação caractere por caractere em busca de padrões. Singla e Garg (2012) relataram diversas técnicas de casamento de padrões e sinalizaram que a técnica Rabin-Karp Algorithm (KARP e RABIN, 1987) é eficiente em localizar vários conjuntos de padrões em um determinado texto. Segundo os autores, o algoritmo Rabin-Karp é geralmente usado para localizar plágios em documentos.

2.6.2 Algoritmo Rabin-Karp

O algoritmo evita fazer comparações caractere por caractere utilizando uma janela de busca. Cada padrão da lista de padrões e as janelas de buscas têm um tamanho calculado. O algoritmo, antes de comparar as janelas (padrão e busca), executa uma verificação numérica entre o padrão e a janela. Se o valor da janela não for o mesmo do padrão, não há a necessidade de verificar os caracteres, pois o valor já não permite a ocorrência de casamento (CHOUDHARY, ASHAR e KULKARNI, 2006; KARP e RABIN, 1987; MOREIRA, 2012).

A utilização das janelas permite mapear sequências de caracteres em um valor numérico. Estes valores numéricos são comparados primeiramente antes de analisar caracteres individuais. Por exemplo: suponha-se que um determinado padrão tem o tamanho da janela 5 e a cadeia a ser comparada do texto tem o tamanho da janela 7. Pode-se afirmar, assim, que o padrão e a cadeia não são iguais, descartando a utilização de comparação caractere por caractere (CHOUDHARY, ASHAR e KULKARNI, 2006; KARP e RABIN, 1987; MOREIRA, 2012). A Figura 2-4 ilustra como é feito o cálculo da janela para ocorrências de padrões. Para a ilustração, considere a=1, b=2, c=3 e d=4:

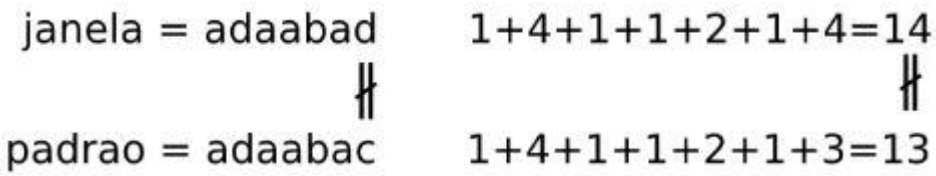

Figura 2-4: Exemplo do cálculo da janela para ocorrência de padrão (MOREIRA, 2012).

Conforme ilustrado na Figura 2-4, pode-se verificar que a soma dos valores dos caracteres entre a janela e o padrão são diferentes. Desta forma, o algoritmo já descarta a janela e prossegue para a próxima cadeia. Já a Figura 2-5 ilustra uma ocorrência de padrão, na qual a janela e o padrão têm o mesmo tamanho.

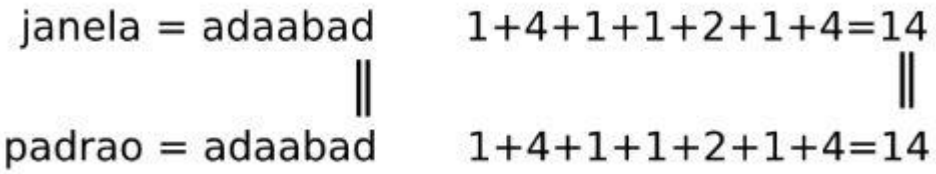

Figura 2-5: Exemplo do cálculo da janela para ocorrência de padrão (MOREIRA, 2012).

Conforme ilustrado na Figura 2-5, ocorre que o tamanho da janela é o mesmo do padrão. Desta forma, o algoritmo compara caractere por caractere, pois pode ocorrer que a somatória dos valores dos caracteres sejam iguais, porém com caracteres diferentes. Após verificação caractere por caractere, o algoritmo confirma se houve um casamento ou não. Para tentar evitar as comparações caractere por caractere, foram implementas otimizações no algoritmo original, incluindo, por exemplo, a utilização de função *hash* (MOREIRA, 2012; SINGLA e GARG, 2012).

2.7 Considerações finais

Neste capítulo foram abordados conceitos sobre políticas de privacidade e fundamentos das técnicas utilizadas no desenvolvimento deste trabalho. Foram relatadas breves considerações sobre as técnicas TF-IDF, LDA, Skip-gram e os

algoritmos Boyer-Moore e Rabin-Karp, com a finalidade de situar o leitor sobre algumas técnicas existentes, pois o objetivo deste trabalho não é detalhar as técnicas de mineração de dados e sim reutilizar as implementações disponíveis para geração de Rótulo de Privacidade.

Capítulo 3

TRABALHOS RELACIONADOS

Diversos trabalhos têm sido propostos pela comunidade acadêmica e estão relacionados com o tema abordado nesta pesquisa. Este capítulo contextualiza estes trabalhos, descrevendo as características daqueles que possuem maior relação com o desenvolvimento deste estudo.

3.1 Considerações iniciais

Considerando o objetivo de extrair informações de políticas de privacidade escritas em linguagem natural, buscou-se na literatura trabalhos em que o Processamento de Linguagem Natural (PLN) é focado para análises de políticas de privacidade. Foram encontrados diversos trabalhos, porém selecionados somente os que contribuíram direta e indiretamente para o desenvolvimento deste estudo.

A subseção 3.2 apresenta o principal trabalho relacionado com esta proposta, tendo sido a motivação principal para a utilização de uma tabela para apresentação dos termos e propósitos de privacidade. A subseção 3.3 descreve trabalhos que utilizam técnicas para analisar políticas textuais e outros que definem taxonomias para as políticas.

3.2 Rótulo de Privacidade

Kelley et al. (2009) utilizaram-se da ideia de rótulos nutricionais, aqueles que são impressos nas embalagens de produtos alimentícios.

A proposta final fundamentou-se em uma tabela similar às tabelas nutricionais, porém em formato de matriz, sendo que, na primeira linha, apresenta os tipos de serviços pelos quais as informações são coletadas. Na primeira coluna tem-

se o tipo de informação que pode ser coletada. No centro da matriz há regiões demarcadas em cores, representando um indicativo de qual tipo de informação está sendo coletada e se o usuário pode ou não intervir na coleta das informações, conforme ilustrado na Figura 3-1. Para geração da tabela, usou-se a plataforma P3P.

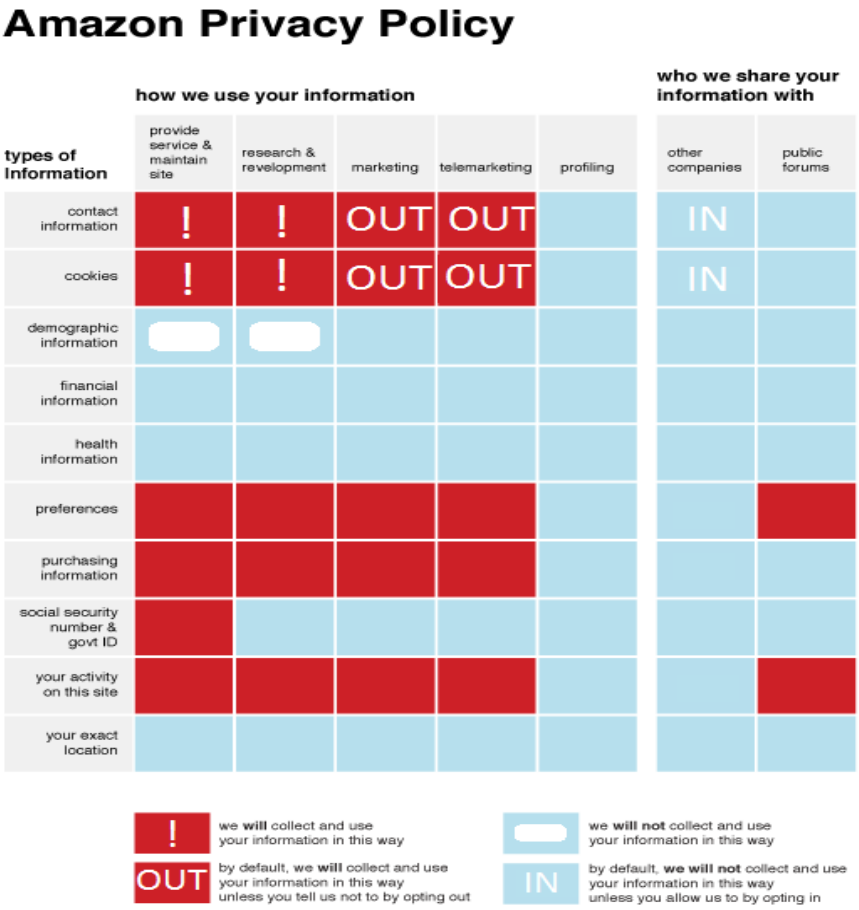

Figura 3-1: Rótulo Nutricional adaptado por Kelley et al. (2009)

A tabela da Figura 3-1 contém alguns símbolos que necessitam de legendas de entendimento, apresentadas na Tabela 3-1.

Kelley et al. (2009) fizeram um experimento para validar a utilização do Rótulo de Privacidade proposto com 24 participantes, sendo 16 estudantes e oito não estudantes. Na condução do experimento pelos autores, foram utilizados dois tipos de apresentações de políticas de privacidade, sendo uma apresentada em linguagem natural e uma no formato de tabela.

Tabela 3-1: Significado dos símbolos da legenda da tabela elaborada por Kelley et al. (2009)

Símbolo	Descrição
Ponto de exclamação	Dados do usuário são coletados e utilizados para a respectiva posição na tabela.
Out	Usuário pode optar em não fornecer os dados, ele tem a opção de não divulgar.
In	Dados não são coletados como padrão, mas tem a opção de escolher para divulgar tais informações.
Quadrado com retângulo	Usuário tem a opção de escolha mista. Ele pode tanto optar por fornecer ou não fornecer alguns dados.

Com os resultados da avaliação das apresentações das políticas de privacidade, os autores afirmaram que a utilização da tabela permite que a informação de coleta de dados pode ser encontrada no mesmo lugar o tempo todo, facilitando a rápida visualização pelo leitor. O rótulo também remove as terminologias complicadas. São utilizados somente quatro símbolos para a legenda da tabela.

Os participantes do experimento que utilizaram as políticas de privacidade em tabela classificaram o seu emprego como prazeroso. Não somente foram classificadas melhor do que as políticas em linguagem natural, mas também ficaram mais agradáveis de ler.

3.3 Mecanismos para análise de políticas e definições de taxonomias

Adams et al. (2010) desenvolveram um mecanismo de auditoria de políticas de privacidade. O mecanismo tem como foco monitorar as políticas de privacidade dos serviços on-line de que os usuários fazem uso. Ele informa ao usuário se houve

mudanças na política de privacidade, sendo fácil de visualizar as alterações ocorridas.

O sistema proposto pelos autores é dividido em três componentes: (i) Monitor de Política: onde o mecanismo monitora ativamente as políticas dos sites que o usuário utiliza ou já utilizou; (ii) Biblioteca de política: é o repositório de armazenamento de políticas ativas e suas versões anteriores, caso houver; (iii) Cliente de auditoria de políticas: é a ferramenta que fornece aos usuários finais informações e notificações das políticas de privacidade, tais como localização da política de privacidade no site, se o usuário já visitou tal página e se visitou, informa se houve alteração no texto da política.

O protótipo do mecanismo funciona da seguinte forma com a utilização dos três componentes: o mecanismo acessa a biblioteca de políticas para verificar se o monitor de política já analisou o serviço que está sendo utilizado. Se já houve visita no site, o cliente de auditoria informa ao usuário se houve alterações no texto, se o usuário leu a política. Caso seja a primeira vez que o monitor acessa o serviço, a política é armazenada na biblioteca de políticas e disponibiliza um botão no aplicativo para o usuário acessar facilmente a página na qual está hospedado o texto da política.

Segundo Adams et al. (2010), os principais benefícios da utilização do mecanismo são: alertar os usuários de uma maneira fácil sobre a existência de políticas, poupando-os de procurar tais textos nos serviços que utiliza; indicar a existência de políticas imediatamente após visitar o site; e alertar os usuários quando as políticas de privacidade forem alteradas, fazendo com que eles não tenham de se preocupar em ficar verificando se os textos foram alterados.

A Figura 3-2 ilustra a forma como as alterações das políticas de privacidade são apresentadas para os usuários.

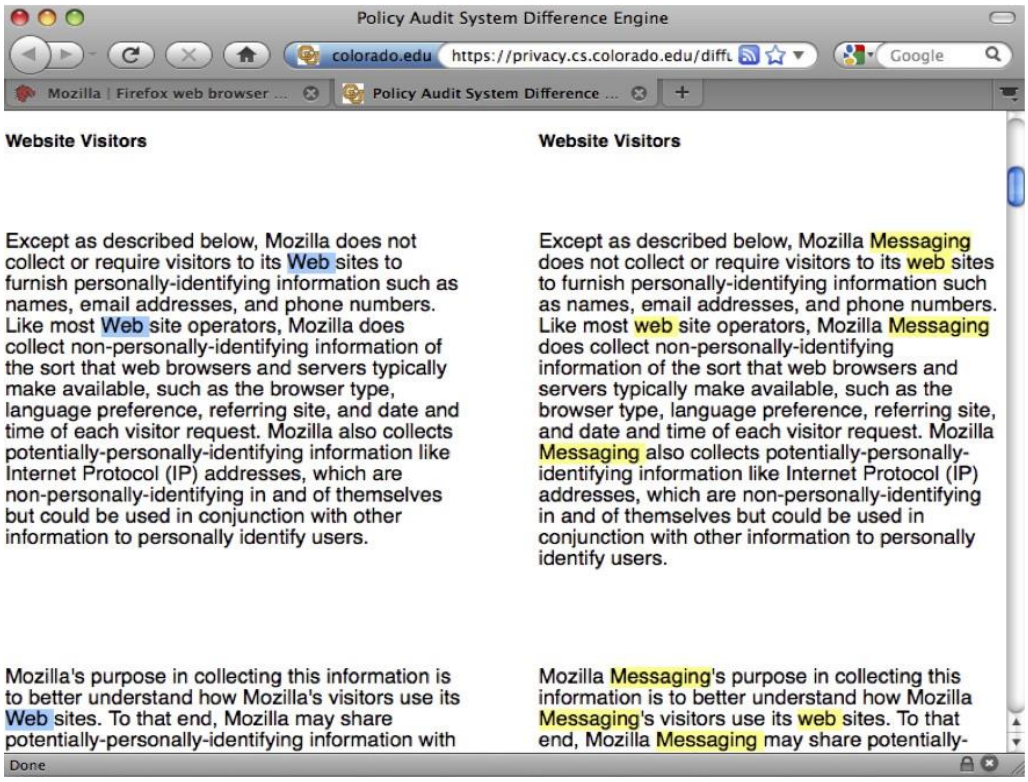

Figura 3-2: Interface do mecanismo do monitoramento de políticas (Adams et al., 2010)

Conforme ilustrado na Figura 3-2, as alterações que houver nas políticas ao longo do tempo que o usuário utiliza um determinado serviço on-line são marcadas com cores diferentes. No lado esquerdo há o texto da política que estava armazenado na biblioteca de políticas e do lado direito é apresentado o novo texto. O que está marcado na cor azul são os termos antes das alterações e a marcação em amarelo mostra o que foi alterado.

No desenvolvimento de software, quando os desenvolvedores estão descrevendo os requisitos do software, a documentação é elaborada em linguagem natural. Na documentação do software existem as Políticas de Controle de Acesso (PCAs). Uma PCA descreve direitos de acesso de alto nível que determinam se um usuário pode ou não acessar algum tipo de informação que é diferente do cargo que ele exerce dentro do software (Ferraiolo et al., 2001).

Para auxiliar os engenheiros de requisitos, Xiao et al. (Xiao, Paradkar e Xie, 2011) desenvolveram um aplicativo chamado *Text2 Policy*. O foco da proposta dos autores é extrair automaticamente PCAs de documentos de softwares e apresentar as PCAs para os engenheiros de requisitos, sem que seja necessária uma análise manual na documentação.

O *Text2 Policy* utiliza dois modelos, sendo um modelo de PCA e um modelo *step-action*. Modelo PCA são textos escritos em linguagem natural e o modelo *step-action* são regras que determinam o acesso de um usuário no sistema. Por exemplo: quando o usuário fizer o login no sistema e tentar acessar alguma informação, é verificado se esse usuário tem autorização para acessar. Caso tenha, o acesso é liberado, do contrário é bloqueado. Essas regras são definidas para diminuir os riscos de privacidade, e, no caso da proposta dos autores, estão voltadas para a área da saúde.

O processo de extração automática de PCAs nos documentos de software pelo *Text2 Policy* executa três passos: (i) aplica análise linguística para fazer um *parser* dos documentos em linguagem natural e anotar palavras e frases com significados; (ii) constrói instâncias do modelo usando as palavras e frases que foram anotadas; e (iii) transforma essas instâncias do modelo em especificações formais.

O *parser* utilizado pelos autores para analisar e anotar as palavras e frases com significados foi o *shallow-parsing* proposto e desenvolvido por Neff, Byrd e Boguraev (2003). O *parser* anota as sentenças em frases, cláusulas e funções gramaticais.

Para auxiliar o *parser* o *Text2 Policy* utiliza um dicionário de domínio para associar verbos com classes semânticas predefinidas. Os autores elaboraram vários padrões de semântica na função gramatical das frases feita pelo *parser* e com isto é aplicado um casamento de padrões semânticos para identificar se uma sentença é uma sentença de PCA. A Figura 3-3 ilustra os padrões semânticos usados na abordagem.

Semantic Pattern	Examples
Modal Verb in Main Verb Group	An HCP [subject] **can view** [action] the patient's account [resource]. An admin [subject] **should not update** [action] patient's account [resource].
Passive Voice followed by To-infinitive Phrase	An HCP [subject] **is disallowed to update** [action] patient's account [resource]. An HCP [subject] **is allowed to view** [action] patient's account [resource].
Access Expression	An HCP [subject] **has read** [action] **access to** patient's account [resource]. A patient's account [resouce] **is accessible** [action] **to** an HCP [subject].
Ability Expression	An HCP [subject] **is able to read** [action] patient's account [resource]. An HCP [subject] **has the ability to read** [action] patient's account [resource].

Figura 3-3: Padrões utilizados para combinar os padrões semânticos (XIAO, PARADKAR e XIE, 2011)

Conforme ilustrado na Figura 3-3, o texto em negrito representa uma parte de uma frase que corresponde a um determinado padrão semântico. Com a utilização desses padrões semânticos, a abordagem dos autores filtra sentenças em documentos escritos em linguagem natural que não combinam com os padrões que foram fornecidos. Quando o mecanismo for analisar uma frase, ele verifica o sujeito e o objeto, sendo que o sujeito é o usuário que acessa os recursos do sistema e o objeto é um recurso definido do sistema.

O *Text2 Policy* tem como propósito controlar a privacidade de pacientes, relacionando o nível de acesso dos funcionários com o que está descrito nos documentos de PCA.

Costante et al. (2012) relatam a dificuldade do entendimento dos usuários para com as políticas de privacidade. Descrevem que o controle do usuário em negociar sua privacidade é quase limitado. Com o problema da compreensão das políticas de privacidade, os autores desenvolveram um mecanismo para avaliar a integridade das políticas de privacidade dos serviços utilizados pelos usuários.

A avaliação da integridade das políticas de privacidade é feita por meio de um conjunto de configurações de privacidade fornecido pelo usuário, tal como preferência em compartilhar localização, dados de navegação, dados pessoais. O mecanismo verifica no texto de política se os termos descritos estão de acordo com as preferências definidas pelo usuário e é fornecida a ele uma notificação. A verificação no texto da política é feita por meio da utilização de técnicas de classificadores de aprendizagem de máquina.

O analisador de integridade tem como principal objetivo analisar a integridade dada uma política de privacidade. Os autores elaboraram uma equação para calcular a integridade da seguinte forma:

$$Gc(p) = N.\sum_{i=1}^{n} w_{i.}c_i$$

onde $N = \frac{10}{\sum_{i=1}^{n} w_i}$ é um fator de normalização[9] que comporta valores entre 0 e 10, n é o número de categoria de privacidade, w_i é o peso para a categoria informado pelo usuário e c_i é o nível de cobertura, tendo os valores de 1 para

[9] Se todos os pesos w_i são 0, ou seja, se o usuário não informou nenhuma preferência para categoria i, o valor definido para a categoria é 10.

coberto e 0 para não coberto. Dessa forma, dada uma política de privacidade, o analisador de integridade faz a análise e classifica cada um dos parágrafos do texto da política. A visão geral da integridade é ilustrada na Figura 3-4, onde a classificação geral da integridade da política é definida em cores. As categorias que são cobertas são representadas em verde, não cobertas são apresentadas em vermelho e não relevantes são apresentadas em cinza.

Para calcular o valor de c_i, foram aplicadas classificação de texto e técnicas de aprendizagem de máquina. Para a classificação de texto foi feito um treinamento a partir de um conjunto de documentos pré-classificados, no qual o conjunto de dados para o treinamento são parágrafos retirados de um corpus de políticas de privacidade, contendo 64 textos e cada parágrafo dos textos foram demarcados manualmente, resultando em 1049 parágrafos, utilizando 772 para o treinamento e 277 para a validação do treinamento.

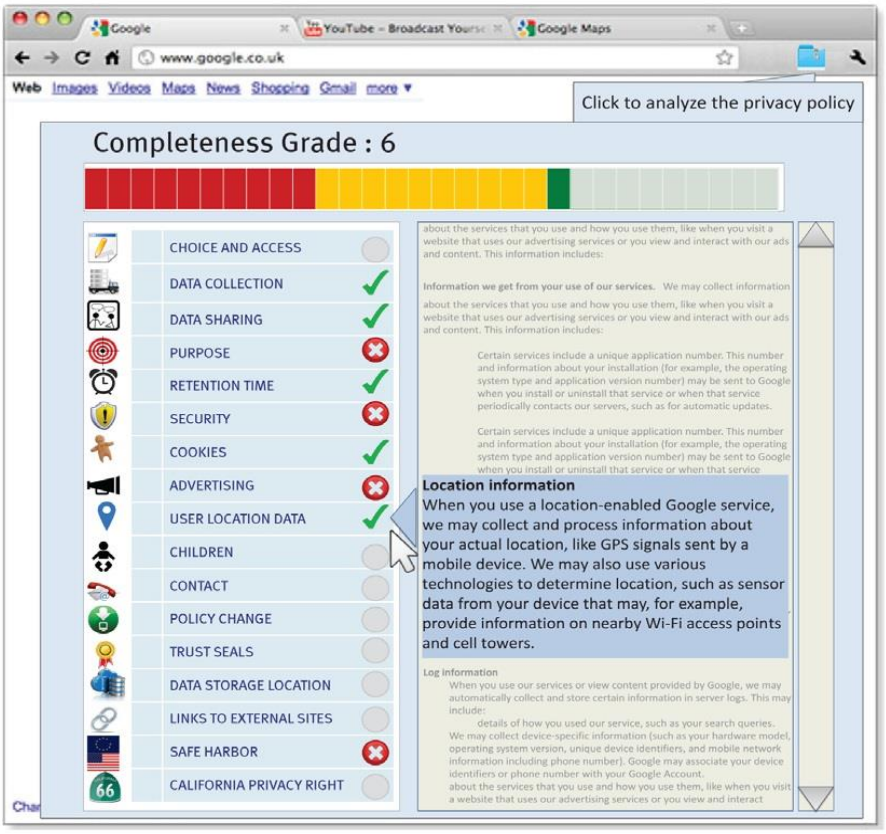

Figura 3-4: A interface do analisador de integridade (Costante et al., 2012)

Com os parágrafos demarcados, cada um teve uma classificação de categoria, que foi definida em leis de organizações que ditam regras e diretrizes relacionadas com privacidade, tais como a OECD e *Children's Online Privacy*

Protection[10], que é uma subseção de diretrizes da FTC. Exemplos de categorias de privacidade podem ser Coleta de Dados ou Compartilhamento de Dados.

Quando o mecanismo for avaliar uma política de privacidade a pedido do usuário, o mecanismo irá processar o texto e aplicará o classificador já treinado no corpo do texto, apresentando, assim, a integridade da política para com as preferências do usuário.

Quando vão projetar uma nova biblioteca, desenvolvedores de software descrevem as funcionalidades da biblioteca em texto de linguagem natural, que são as documentações que acompanham as *Application Programming Interface* (API).

A utilização de métodos, tipos de retornos, precondições, pós-condições, limitações e mais informações da API que são relevantes para os desenvolvedores estão especificadas nas documentações da biblioteca. Tais especificações podem ser valiosas para ferramentas automatizadas que são destinadas à melhoria da produtividade dos desenvolvedores e/ou testadores de software. Porém, essas ferramentas não foram projetadas para trabalhar com linguagem natural, sendo necessário ter uma linguagem formal definida para conseguir automatizar o processo de desenvolvimento ou de testes. Dessa forma, há um conflito entre as entradas que são exigidas por essas ferramentas e a documentação que foi elaborada pelo desenvolvedor. Diante desse conflito, Pandita (2013) propôs o framework *WHYPER*, no qual é utilizado PLN, tendo a finalidade de verificar, nas documentações dos softwares, se é necessária alguma permissão para acessar algum tipo de dado.

O *framework* foi definido pelos autores contendo um analisador, um pré-processador, um motor de análise de texto, um pós-processador e um gerador de contrato de código, onde: (i) o analisador aceita os documentos de API e extrai o conteúdo das descrições dos métodos; (ii) o pré-processador altera as sentenças em uma representação formal com alguns metadados; (iii) o motor de análise de texto recebe a representação formal das frases e com modelos semânticos uma nova especificação é gerada em expressões de Lógica de Primeira Ordem (LPO); (iv) o gerador de contrato de código recebe as expressões LPO e gera uma relação de mapeamento das expressões para a construção dos códigos na linguagem de programação-alvo.

[10]https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

Para a elaboração deste trabalho, o componente mais importante da proposta de Pandita (2013) é o de motor de análise de texto, que recebe as sentenças já pré-processadas e constrói a expressão LPO. A construção é dada da seguinte forma: primeiramente, a frase LPO gerada é marcada com algumas tags de LPO. Depois de feitas as marcações, a técnica de PLN, chamada de análise superficial (WAIN et al., 2012), é aplicada nas frases. Um analisador de análise superficial aceita *tokens* lexicais gerados pela LPO alvo e faz tentativas de classificar as sentenças com base em modelos semânticos predefinidos. Os modelos semânticos são ilustrados na Figura 3-5.

	Name	Example	Description
1.	Predicate (Name)	The (path)$_{subject}$ (can not be)$_{verb}$ null$_{object}$	The subject and object form the terms of the predicate represented by verb.
2.	Conditional followed or preceded nominal predicate	If (path does not have extension)$_{conditional}$, (GetExtension)$_{subject}$ (returns)$_{verb}$ (System.String.Empty)$_{object}$	The subject-verb-object forms specification as described in row 1, which is true when the condition highlighted by $conditional$ is true. The condition is further resolved using one of the templates.
3.	Prepositional predicate	(Path)$_{subject}$ (is)$_{verb}$ (not null or empty String)$_{preposition}$	The verb forms the partial predicate and the subject forms one of the terms. The second term and the remaining of the predicate are extracted by resolving the preposition.
4.	Transitive predicate	(Name)$_{subject}$ (is)$_{verb}$ a (valid , identifier)$_{object-subject}$, which (is no longer than 32 characters)$_{clause}$	The sentence is broken down into two sentences. The first sentence ends with the phrase labeled $_{object-subject}$, and the second sentence begins with the phrase labeled $_{object-subject}$. Each sentence is further resolved and the resulting specifications are joined using the logical AND operator.

Figura 3-5: Categorias de modelos semânticos (PANDITA, 2013)

Conforme ilustrado na Figura 3-5, a coluna descrição diz o que é inferido a partir do padrão semântico aplicado. Um exemplo: para um modelo descrito na primeira linha da Figura 3-5, a expressão de LPO é construída como "não pode ser *(path, null)*, onde *"path"* e *"null"* são termos para o predicado "não pode ser". A especificação interpreta o predicado e avalia se os termos são verdadeiros ou não. Outro exemplo foi explorado pelo autor: o motor de análise de texto usa o padrão semântico predicado transitivo, descrito na quarta linha da Figura 3-5. Com esse padrão o motor analisa a frase que está descrita na linha 3 da Figura 3-6.

```
01:/// <summary>
02: .....
03:/// <param name=''prop_name''> This name
    needs to be a valid identifier, which
    is no longer than 32 characters, starting
    with a letter (a-z) and consisting of only
    small letters (a-z) numbers (0-9), and/or
    underscores.</param>
04: .....
05:public void DefineObjectProperty(string
    obj_type, string prop_name,
       int prop_type)
```

Figura 3-6: Frase exemplo para reconhecimento do padrão de análise (PANDITA, 2013)

Após a análise executada pelo motor de análise de texto, uma expressão LPO gráfica é gerada, conforme é apresentado na Figura 3-7.

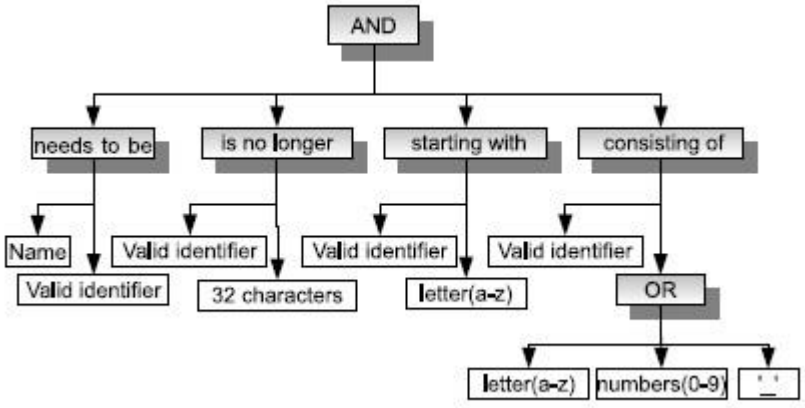

Figura 3-7: Especificação do formato das expressões de Lógica de Primeira Ordem (Pandita, 2013)

De acordo com a Figura 3-7, cada nó interno (sombreado de cinza) representa um predicado e os filhos desses nós representam os termos para este predicado. Em seu trabalho, Pandita (2013) fez um experimento com o *WHYPER* para analisar três tipos de permissões que a maioria de aplicações requerem, sendo elas: catálogo de endereços, calendário e gravador de áudio. Os resultados mostraram que *WHYPER* identifica efetivamente as frases que descrevem as necessidades de tais permissões com precisão média de 82,8%.

Massey et al. (2013) desenvolveram um modelo para contemplar três propósitos: (i) avaliar a capacidade de leitura de documentos onde constam regras para privacidade para engenheiros de requisitos; (ii) determinar se a mineração de

texto automático pode indicar se um documento de software contém requisitos que expressem dizeres para proteção de privacidade e vulnerabilidade; (iii) estabelecer uma generalização do item (ii) na identificação das proteções de privacidade e vulnerabilidades em novos documentos.

Para a elaboração do modelo, primeiramente os autores recolheram vários conjuntos de documentos de política para examinar. Os documentos foram retirados de requisitos de softwares baseados em compromisso com privacidade, os tops sites elencados pelo Google e documentos das maiores empresas dos Estados Unidos disponibilizados pelo site Fortune[11].

Os autores utilizaram o modelo de tópicos probabilístico proposto por Blei, Carin e Dunson (2010) para analisar, anotar as palavras dos textos e separá-las em temas que sejam similares. O modelo tem como objetivo encontrar similaridades entre as palavras, como, por exemplo, as palavras "saúde", "hospitais" e "medicina", todas relacionadas a um tema similar.

O esboço do método proposto pelos autores prevê: (i) pré-processar os documentos de políticas selecionados; (ii) selecionar um subconjunto de dados para utilizar na validação do modelo; (iii) criar uma série de modelos de tópicos; (iv) realizar uma melhor validação nos ajustes dos dados selecionados para a validação do modelo, determinando qual o melhor modelo para ser utilizado na análise de requisitos; e (v) determinar uma medida em que o modelo pode ajudar a produtividade na engenharia de requisitos.

Para a construção de um modelo de tópico são necessárias duas entradas: (i) um corpus de documentos; e (ii) um número de tópicos definidos para compor o corpus. Para construir o modelo de temas, os autores reuniram todos os documentos do corpus em um único documento.

Para encontrar os temas em um novo documento é aplicada a técnica *Latent Dirichlet Allocation* (LDA), desenvolvida por Blei, NG e Jordan (2003). A técnica aplica cálculos probabilísticos para encontrar o modelo de temas predefinidos em novos documentos a serem analisados.

Os autores desenvolveram o modelo descrito para abordar preocupações que são de interesse de elaboradores de políticas e de indivíduos relacionados com políticas, verificando se elas contêm proteções preestabelecidas ou vulnerabilidades

[11] http://fortune.com/

implícitas. Tanto os elaboradores quanto os indivíduos precisam de ferramentas para realizar essa avaliação em um grande conjunto de documentos, diminuindo o tempo de leitura dos documentos, auxiliando os engenheiros de requisitos na redução das vulnerabilidades e verificando as proteções focadas em privacidade.

3.4 Considerações finais

Os trabalhos relacionados apresentam algumas soluções para extrair informações em textos de políticas de privacidade de forma automática.

O trabalho desenvolvido por Kelley et al. (2009) apresenta uma importante contribuição em relação à apresentação das políticas de privacidade, sendo estas políticas visualizadas em formato de tabela. Os autores utilizaram as categorias de privacidade descritas nos arquivos da plataforma P3P e com estas categorias a tabela era gerada automaticamente. A necessidade da utilização da plataforma P3P torna a proposta restrita, pois a maioria dos serviços on-line utiliza a linguagem natural para descrever suas políticas de privacidade. A concepção de Tabela de Privacidade deste trabalho foi a motivação para a presente proposta, mas sem ter a necessidade de que a política de privacidade esteja descrita no formato P3P.

O trabalho proposto por Adams et al. (2010) apresenta um sistema para monitorar as políticas de privacidade dos websites das quais o usuário faz uso. Quando o usuário acessa um website, o aplicativo coleta a política e a compara com a sua biblioteca de políticas. Caso o usuário já tenha acessado, o sistema compara as políticas referentes aos acessos anteriores com o atual. Caso tenham ocorrido alterações no decorrer do tempo, o sistema alerta o usuário de que os termos de políticas podem ter sido alterados e assim o notifica, fazendo uma comparação entre as duas políticas e marcando as alterações encontradas. A proposta de comparação das políticas de privacidade elaborada pelos autores objetiva em verificar se houve alterações nos termos das políticas, caso houver, a aplicação faz marcações nos textos ilustrando as diferenças da comparação. Essa comparação elaborada pelos autores motivou a utilização de casamento de padrões no desenvolvimento dessa pesquisa.

Os trabalhos de Xiao et al. (2011) e Pandita (2013) apresentam uma análise de documentos de software escritos em linguagem natural utilizando-se de análise linguística. As técnicas abordadas pelos autores são semelhantes, sendo que o trabalho de Xiao et al. (2011) faz a extração das informações dos documentos utilizando padrões semânticos já definidos e o trabalho de Pandita (2013) faz a extração das informações utilizando expressões de lógica de primeira ordem, que também já são definidas previamente. A concepção desses dois trabalhos motivou a utilização de um formalismo para casamento de padrões, sendo que no trabalho dos autores, caso ocorra um padrão semântico, afirma-se que o texto analisado contém uma política de controle de acesso ou informações sobre privacidade.

O trabalho proposto por Costante et al. traz uma grande contribuição para as análises de políticas de privacidade. Os autores definiram categorias de privacidade nas quais o usuário pode configurar um aplicativo. Quando o aplicativo faz a análise de uma política, o sistema compara as configurações do usuário com os termos descritos nas políticas. Caso a configuração do usuário não seja condizente com os termos do website, o aplicativo o alerta sobre a integridade da política para com as preferências definidas. Esse trabalho foi uma motivação para a utilização de categorias de privacidade, sendo que cada categoria representa alguma ação sobre os dados dos usuários.

Massey et al. desenvolveram um modelo para encontrar similaridades entre as palavras com utilização da técnica LDA. A proposta dos autores contém três propósitos principais: (i) avaliar a capacidade de leitura de documentos onde constam regras de privacidade para engenheiros de requisitos; (ii) determinar se a mineração de texto automático pode indicar se um documento de software contém requisitos que expressem dizeres para proteção de privacidade e vulnerabilidade; (iii) estabelecer uma generalização do item (ii) identificação das proteções de privacidade e vulnerabilidades em novos documentos. Por fim, esse trabalho motivou a utilização de técnicas para encontrar palavras-chaves em um conjunto de documentos e agrupamento de acordo com suas características genéricas, ou seja, de acordo com suas taxonomias.

A Tabela 3-2 apresenta um resumo das contribuições dos trabalhos relacionados no desenvolvimento desta pesquisa.

Tabela 3-2: Contribuições dos trabalhos relacionados

Trabalho	Aplicado em	Contribuições
Kelley et al. (2009)	Políticas de Privacidade de websites	- Motivação da utilização e adaptação da Tabela de Privacidade
Adams et al. (2010)	Documentação de software	- Motivação para utilização de técnicas para casamento de padrões
Xiao et al. (2011) e Pandita (2013)	Documentação de software	- Motivação para a elaboração de um formalismo para ocorrências de casamento de padrões
Costante (2012)	Políticas de Privacidade de websites	- Motivação para a elaboração de categorias de privacidade
Massey et al. (2013)	Documentação de software	- Motivação para utilização de técnicas para encontrar as palavras-chaves e agrupamento de acordo com suas características genéricas (taxonomias)

Capítulo 4

GERAÇÃO DE RÓTULO DE PRIVACIDADE

Este capítulo apresenta uma metodologia para gerar Rótulos de Privacidade para as políticas de privacidade de serviços on-line, permitindo aos usuários terem uma melhor compreensão de quais dados podem ser coletados e para que são utilizados.

4.1 Considerações iniciais

A proposta deste trabalho pressupõe que usuários podem ter dificuldades na compreensão dos termos das políticas de privacidade e também não são todos que estão predispostos a gastar tempo na sua leitura. Também é pressuposto que as políticas de privacidade não seguem um padrão de escrita: às vezes são várias páginas de texto com jargões técnicos, dificultando a determinação do usuário em ler toda a política. Podem, porém, conter palavras-chaves que são comuns entre elas.

Este trabalho apresenta uma metodologia para extrair palavras-chaves dos textos das políticas, criando categorias de privacidade sobre as informações da utilização dos dados dos usuários e apresentando-as em um formato de tabela, que denominou-se de Rótulo de Privacidade.

Para o desenvolvimento deste trabalho foi necessário um conjunto de documentos de políticas de privacidade, que será descrito na seção seguinte e nas subseções 4.3.1 e 4.3.2. São apresentados exemplos da execução das fases da metodologia proposta e a forma como foram utilizados para a avaliação deste trabalho.

4.2 Dados requeridos pela abordagem

Para a elaboração da abordagem proposta e posteriormente sua avaliação, as fases descritas a seguir foram sobre um conjunto de documentos de políticas de privacidade. Este conjunto de documentos será utilizado para encontrar as palavras mais comuns nos textos e para a avaliação da abordagem. Para a criação do conjunto de políticas foram selecionados 60 sites mais acessados no Brasil, segundo o site Alexa[12]. Este site elenca os mais acessados de vários países. O tamanho definido da amostra como sendo de 60 sites foi fundamentado nas diretivas de Sardinha (2000), que define critérios para a seleção de corpus. Os critérios abordados pelo autor abrangem desde o modo do corpus, tal como falado ou escrito, até a finalidade de sua utilização.

[12]http://www.alexa.com/topsites/countries;0/BR

Dessa forma, a amostra se enquadrou nos seguintes critérios: de seleção, de conteúdo e de finalidade. O critério de seleção foi enquadrado no de amostragem (*sample corpus),* pois é composto por porções de textos e é planejado para ser uma amostra finita do contexto como um todo. O critério de conteúdo se enquadrou no especializado, sendo que os textos são de tipos específicos (textos sobre políticas de privacidade). O critério de finalidade compreendeu o critério de treinamento ou teste-construído, permitindo o desenvolvimento de aplicações e ferramentas de análise.

De cada serviço selecionado, as políticas de privacidade foram extraídas do website e posteriormente armazenadas em arquivos individuais de texto comum *.txt*[13].

Após a seleção dos sites para compor o conjunto de políticas, dos 60 sites escolhidos selecionou-se aleatoriamente, por meio da tabela de números aleatórios, 50 políticas de privacidade para compor o corpus de treinamento, que será utilizado na primeira fase da abordagem, descrita a seguir.

4.3 Geração de Rótulo

A metodologia é composta por duas fases, sendo que a primeira produz um catálogo de palavras-chaves para que, na segunda fase, seja capaz de realizar a análise das políticas e apresentar as informações extraídas. As duas fases foram denominadas neste trabalho como sendo Fase de Conhecimento e Fase de Análise, que são descritas a seguir.

A Figura 4-1 ilustra os passos necessários a serem executados na Fase do Conhecimento.

[13]Os arquivos das políticas de privacidade estão disponíveis em: https://www.dropbox.com/sh/1uqg95n8cf8cyha/AABXWoOLNy9mhr-i4Smzb6KKa?dl=0

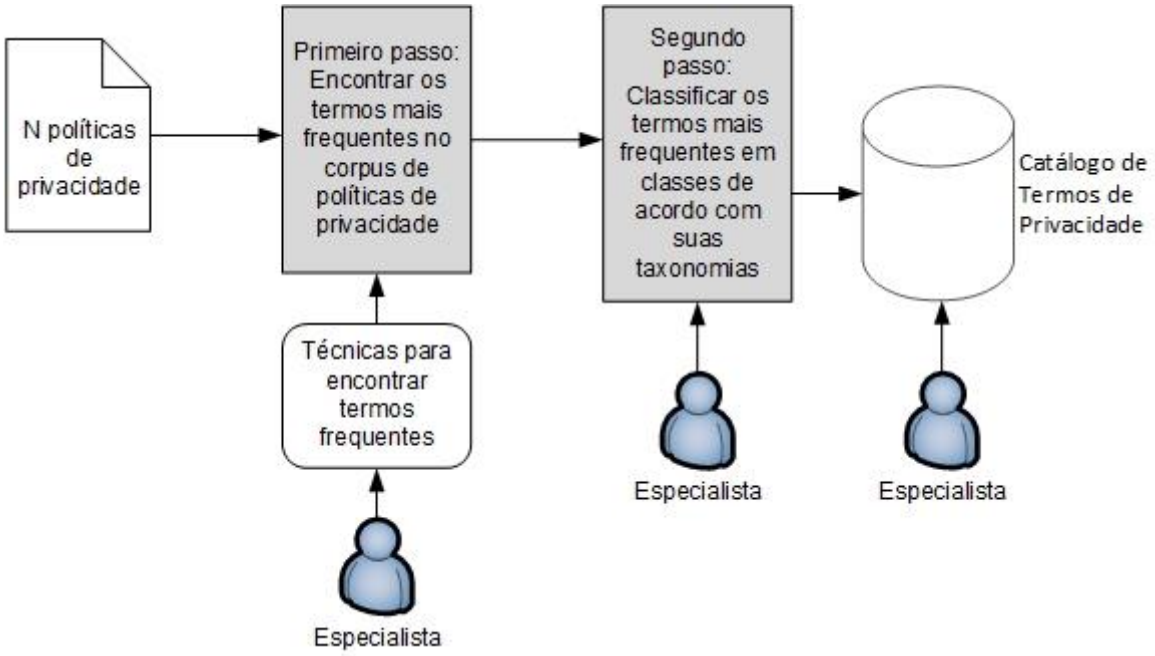

Figura 4-1: Visão geral da Fase de Conhecimento

Para iniciar a fase de conhecimento, de acordo com a Figura 4-1, é necessário um conjunto de políticas de privacidade. Ilustradas como “N políticas de privacidade”, elas foram selecionadas conforme descrito na seção 4.2.

4.3.1 Fase de Conhecimento

A Figura 4-1 ilustra a fase de conhecimento com dois passos: (i) encontrar os termos mais frequentes no corpus de políticas de privacidade e (ii) classificar os termos mais frequentes em classes e categorias relacionados aos seus significados, de acordo com suas taxonomias.

4.3.1.1 Primeiro passo - Encontrar os termos mais frequentes no corpus de políticas de privacidade

Para executar o primeiro passo é necessário utilizar alguma técnica que encontre palavras-chaves ou termos frequentes em um conjunto de textos. A escolha da técnica utilizada para localizar os termos mais frequentes pode ficar a critério do especialista/usuário e, caso ele já possua conhecimento das palavras-chaves de tipos de dados dos usuários que são coletados e para quais finalidades são utilizados nos textos das políticas de privacidade, essa etapa não é necessária.

Os detalhes da execução desse passo são apresentados por um exemplo. Neste exemplo foi utilizada a técnica TF-IDF (RAMOS, EDEN e EDU, 2003), pois, segundo os autores, mostra-se eficiente diante de um pequeno conjunto de documentos e há várias implementações disponíveis.

A técnica TF-IDF faz cálculos das palavras que mais se repetem no conjunto de todos os documentos, fazendo uma avalição matemática sobre os termos, considerando que não são contabilizados os termos que se repetem muitas vezes em apenas um documento, ponderando na distinção das palavras-chaves de todos os documentos.

Para aplicar a técnica sobre o corpus, foi utilizada uma implementação[14] em Java da técnica, que está disponível para a comunidade no repositório *GitHub*[15].A Figura 4-2 ilustra a interface da aplicação para executar a técnica TF-IDF.

De acordo com a Figura 4-2, a aplicação desenvolvida em Java seleciona o diretório em que se encontra o corpus para realizar os cálculos. Deve-se, então, primeiro selecionar o corpus para o processamento da técnica TF-IDF.

O aplicativo executa a técnica por um pré-processamento em todos os arquivos do corpus de políticas. Neste pré-processamento as *stopwords*, advérbios e adjetivos são retirados dos textos e não são contabilizados pela técnica.

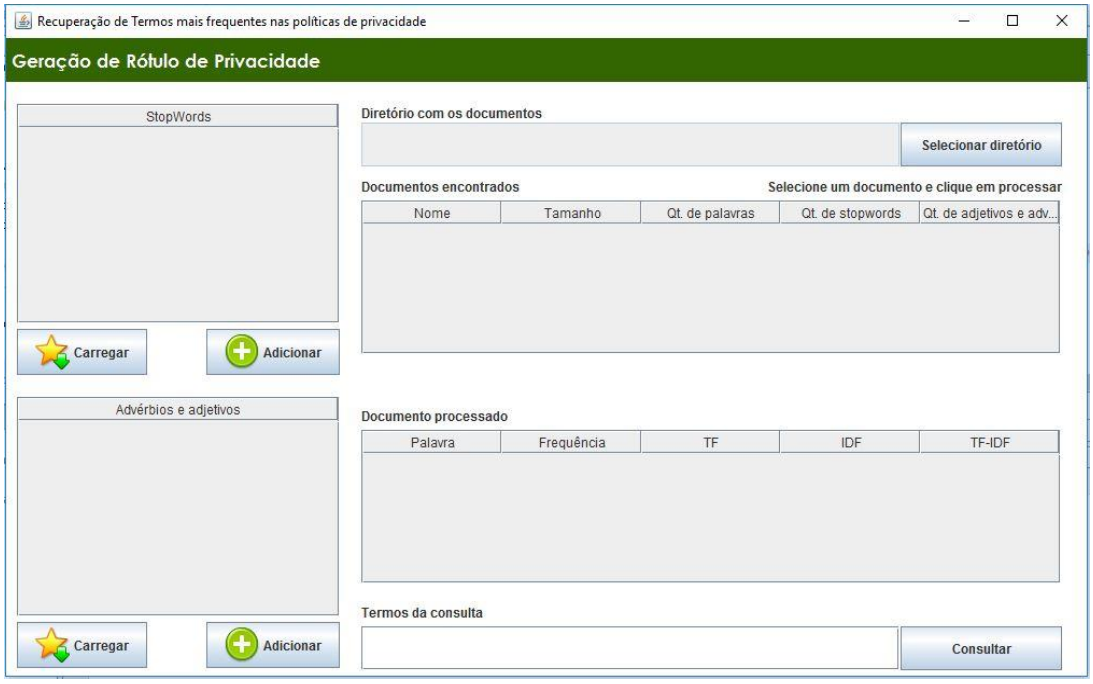

Figura 4-2: Interface aplicação para cálculo do TF-IDF

[14] Disponível em: https://github.com/emmanuelXavier/LSI
[15] Créditos de desenvolvimento para Emmanuel Xavier - membro *GitHub*

A Tabela 4-1 apresenta informações sobre os documentos do corpus que foram analisados para encontrar as palavras-chaves. Também informa o tamanho do arquivo, a quantidade de palavras de cada arquivo, a quantidade de *stopwords* e a quantidade de advérbios e adjetivos encontrados pela aplicação.

A aplicação, além de fornecer informações sobre os arquivos que constituem o corpus, fornece todos os termos relevantes encontrados - sem *stopwords*, advérbios e adjetivos - com o coeficiente TF. Na amostra fornecida obteve-se 20 mil termos com coeficientes entre $0{,}5026$ e $1{,}0000$.

Dentre esses 20 mil termos fornecidos, foi necessário fazer uma filtragem, pois há vários termos que não são relevantes para a presente pesquisa. Para isso definiu-se um valor de corte para o coeficiente TF, sendo a média de todos os coeficientes, resultando no valor de $0{,}5246$.

Após a filtragem dos termos pela média dos coeficientes, foi necessário selecionar os termos para compor o catálogo. Para isso foram definidas duas questões sobre o termo que está sendo analisado (ANTÓN e EARP, 2004), sendo: (i) “É algum termo que pode definir alguma informação sobre o usuário?” e (ii) “É algum termo que isolado ou associado a outro pode definir algum tipo de propósito de utilização da coleta de dados?”. Tais questões são baseadas na caracterização das informações dos usuários definidas por Pearson (2009). Com o valor de corte para o coeficiente e com as questões sobre os termos, foram selecionados 37 termos, que são apresentados na

Tabela 4-2.

Tabela 4-1: Dados sobre os arquivos do corpus de políticas de privacidade

Nome arquivo	Tamanho em bytes	Qt. de palavras	Qt. de stopwords	Qt. de advérbios e adjetivos
politica01.txt	21836.0 bytes	3074	1453	74
politica02.txt	18267.0 bytes	2678	1424	77
politica03.txt	2547.0 bytes	380	183	5
politica04.txt	18358.0 bytes	2672	1297	49
politica05.txt	8747.0 bytes	1273	619	26
politica06.txt	12117.0 bytes	1763	841	21
politica07.txt	19153.0 bytes	2742	1246	99
politica08.txt	23152.0 bytes	3485	1818	78
politica09.txt	34168.0 bytes	5091	2553	189
politica10.txt	3018.0 bytes	436	211	6
politica11.txt	55779.0 bytes	8092	3916	205
politica12.txt	19271.0 bytes	2823	1468	64
politica13.txt	48635.0 bytes	7383	3297	195
politica14.txt	15067.0 bytes	2214	1071	37
politica15.txt	10625.0 bytes	1565	773	66
politica16.txt	3653.0 bytes	542	261	13
politica17.txt	21035.0 bytes	3022	1416	55
politica18.txt	13001.0 bytes	1867	897	41
politica19.txt	1259.0 bytes	183	90	4
politica20.txt	3714.0 bytes	512	230	12
politica21.txt	7813.0 bytes	1111	497	27
politica22.txt	1728.0 bytes	247	116	3
politica23.txt	7376.0 bytes	1197	600	27
politica24.txt	6550.0 bytes	961	456	28
politica25.txt	10238.0 bytes	1472	684	31
politica26.txt	21534.0 bytes	3146	1510	78
politica27.txt	25753.0 bytes	3806	1682	135
politica28.txt	23704.0 bytes	3459	1634	100
politica29.txt	14136.0 bytes	2124	943	60
politica30.txt	4835.0 bytes	701	348	24
politica31.txt	5853.0 bytes	904	459	14
politica32.txt	4218.0 bytes	622	279	13
politica33.txt	16169.0 bytes	2410	1206	80
politica34.txt	10444.0 bytes	1557	733	73
politica35.txt	8089.0 bytes	1241	649	29
politica36.txt	2407.0 bytes	354	162	11
politica37.txt	6956.0 bytes	1035	415	35
politica38.txt	2202.0 bytes	328	160	10
politica39.txt	16067.0 bytes	2364	1108	96
politica40.txt	10489.0 bytes	1519	733	35
politica41.txt	10533.0 bytes	1521	733	35
politica42.txt	5886.0 bytes	909	432	21
politica43.txt	4734.0 bytes	688	339	19
politica44.txt	19279.0 bytes	2687	1199	88
politica45.txt	19730.0 bytes	2911	1528	77
politica46.txt	5209.0 bytes	743	332	23
politica47.txt	1609.0 bytes	236	109	11
politica48.txt	5072.0 bytes	758	358	12
politica49.txt	27768.0 bytes	4115	2017	103
politica50.txt	7945.0 bytes	1171	555	35

Tabela 4-2: Palavras-chaves selecionadas

Termo	Coeficiente
internos	0,5248333
aquisição	0,5248333
prestar	0,5277778
prover	0,5416667
desenvolvimento	0,5454545
compras	0,5546875
propaganda	0,5625000
telefone	0,5652174
atividade	0,5714286
identidade	0,5909091
solicitado	0,5909091
cpf	0,5921053
ip	0,6000000
localização	0,6021505
endereço	0,6041667
compartilhamento	0,6041667
preferências	0,6071429
marketing	0,6111111
parceiros	0,6171875
pesquisas	0,6250000
ofertas	0,6250000
e-mail	0,6250000
traçar	0,6250000
produtos	0,6344086
perfil	0,6645570
navegador	0,6818182
navegação	0,7500000
consumidor	0,7500000
empresas	0,7500000
pessoais	0,7795699
contato	0,7857143
informações	0,8125000
cookies	0,8181818
publicidade	0,8181818
serviços	0,9852941
site	1,0000000
dados	1,0000000

Os termos selecionados pelo especialista para compor as palavras-chaves devem ser armazenados. Neste trabalho o armazenamento foi feito em uma planilha, definido como catálogo de termos, para que posteriormente possam ser classificados e agrupados em suas devidas categorias (XIAO, PARADKAR e XIE, 2011). O procedimento está detalhado no passo seguinte.

4.3.1.2 Segundo passo - Classificar os termos mais frequentes em classes ou categorias relacionados aos seus significados de acordo com suas taxonomias

O segundo passo tem como objetivo classificar os termos do catálogo que foi gerado pelo primeiro. Essa classificação deve ser feita pelo especialista/usuário, baseando-se na taxonomia proposta por Solove (2006), nas metodologias de agrupamento propostas por Kelley et al. (2009) e Xiao et al. (2011) e nos tipos de informações dos usuários que são caracterizados por Pearson (2009).

Foram definidas duas classes, sendo que uma agrupa todos os tipos de informações referentes aos usuários que podem ser coletados e a outra agrupa todos os tipos de propósitos de coleta de informações, criando, assim, categorias para cada conjunto de termos.

Por exemplo: suponha-se as informações dos usuários definidas por Pearson (2009): “e-mail”, “cpf”, “localização”, “endereço”, “ip”, “serviços”, “propaganda” e “ofertas”. Agora, considerando as duas questões definidas no primeiro passo, pode-se agrupar os termos “e-mail”, “cpf”, “localização”, “endereço” e “ip” na classe de tipos de informações referentes aos usuários, pois os termos respondem à primeira questão: “É algum termo que pode definir alguma informação sobre o usuário?”. Já os termos “serviços”, “propaganda” e “ofertas” podem ser agrupados na classe de tipos de propósitos de coleta de informações, sendo que os termos respondem à segunda questão: “É algum termo que isolado ou associado a outro pode definir algum tipo de propósito de utilização da coleta de dados?”.

Após agrupar as duas classes é necessário criar categorias de privacidade para os termos que têm similaridades em sua taxonomia. Utilizando-se do exemplo citado, tem-se os termos “e-mail e ''cpf'' que, de acordo com as taxonomias (SOLOVE, 2006), pode-se criar uma categoria genérica nomeada como "Informações pessoais e contato". Já para os termos "localização", "endereço" e “ip”

pode-se criar uma categoria genérica nomeada como "Informações de localização" (XIAO, PARADKAR e XIE, 2011). Após classificar todos os termos em categorias, deve-se registrar no catálogo todos os termos nas respectivas categorias definidas pelo especialista.

Neste trabalho foram utilizados os termos descritos na

Tabela 4-2 para criar as categorias de privacidade (KELLEY et al., 2009; SOLOVE, 2006; XIAO, PARADKAR e XIE, 2011), agrupando as informações com característica ou propósitos semelhantes sob um nome que representasse sua função na forma mais genérica.

Os tipos de dados dos usuários que podem ser coletados foram agrupados da seguinte forma: (i) combinou-se as informações de contato como telefones, e-mails e documentos de identificação em uma simples categoria denominada *informação pessoal e contato*; (ii) combinou-se informações sobre cookies em uma simples categoria denominada *leitura de cookies*; (iii) combinou-se informações geográficas como endereço, localização e IP em uma simples categoria denominada *informações de localização*; (iv) combinou-se informações como histórico de navegação e preferências em uma simples categoria denominada *preferências de navegação*; (v) combinou-se informações sobre compra de produtos para formar a categoria informações sobre *últimas aquisições on-line*; e (vi) informações de atividades no site foi definida como *informações sobre sua atividade nesse site*.

Após selecionar os termos de tipos de dados dos usuários, separou-se os termos sobre propósito de utilização dos dados coletados em cinco principais: (i) prover o serviço solicitado; (ii) pesquisa e desenvolvimento interno; (iii) ações de marketing; (iv) formação de perfil do consumidor e (v) compartilhamento de informações com empresas parceiras, com o auxílio do especialista.

Em cada categoria de privacidade, tanto para tipo de dados como para propósito de coleta, foram atribuídos conjuntos de termos, fazendo-se, assim, os seus agrupamentos, mostrados na Tabela 4-3.

De acordo com a Tabela 4-3, na primeira coluna estão as categorias de privacidade, na segunda há os termos de cada categoria e por fim, na terceira, tem-se o tipo do termo, se ele é uma informação do usuário ou um propósito de utilização

dos dados. As categorias definidas foram mapeadas para uma tabela, de acordo com o protótipo do Rótulo de Privacidade, ilustrado na Figura 4-3.

Tabela 4-3: Classificação dos termos em categorias

Categorias	Termos	Tipo do termo
Informação pessoal e contato	e-mail, contato, cpf, identidade, telefone, dados, pessoais, informações	Tipos de dados
Leitura de cookies	cookies	Tipos de dados
Informações sobre localização	localização, endereço, ip	Tipos de dados
Preferências de navegação	Preferências, navegação, navegador	Tipos de dados
Últimas aquisições on-line	compras, aquisição, produtos	Tipos de dados
Informações sobre sua atividade no site	atividade, site	Tipos de dados
Prover o serviço solicitado	prestar, prover, serviço, solicitado	Propósito de utilização
Pesquisa e desenvolvimento interno	desenvolvimento, internos, pesquisas	Propósito de utilização
Ações de marketing	marketing, publicidade, propaganda, ofertas	Propósito de utilização
Formação do perfil do consumidor	perfil, consumidor, traçar	Propósito de utilização
Empresas parceiras	compartilhamento, parceiros, empresas	Propósito de utilização

Conforme ilustrado na Figura 4-3, pode-se verificar que os tipos de dados estão localizados na primeira coluna do rótulo e os propósitos de coleta estão situados na primeira linha. Dessa forma, tem-se o tipo de dado que pode ser coletado e qual sua utilização.

Para poder mapear as categorias para o Rótulo de Privacidade e as ocorrências dos padrões definiu-se um formalismo, que é detalhado na subseção seguinte.

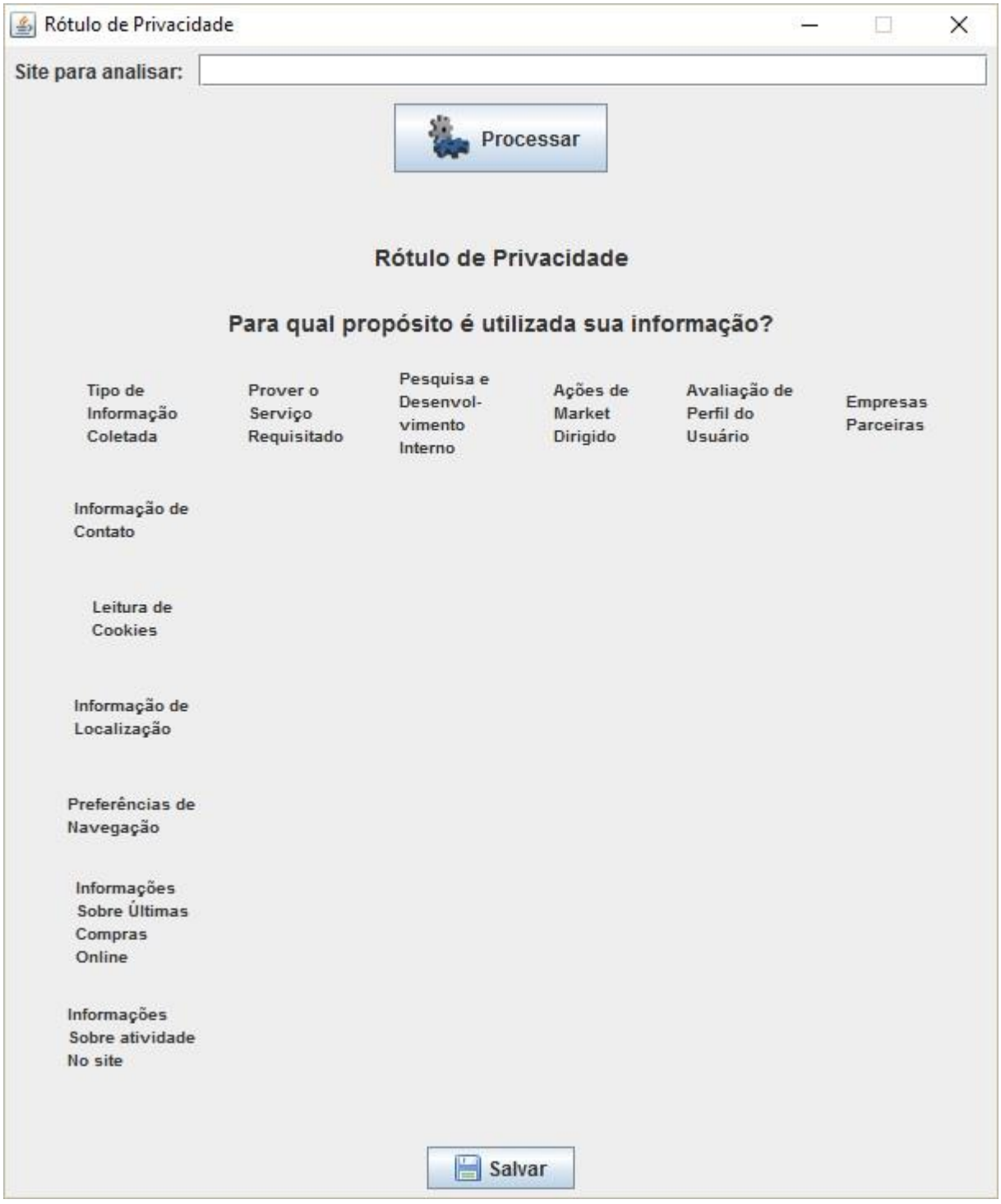

Figura 4-3: Protótipo do Rótulo de Privacidade elaborado pela metodologia

4.3.2 Fase de análise

A fase de análise é responsável por verificar uma particular política de privacidade buscando encontrar ocorrências de casamento de padrões das categorias elaboradas pela primeira fase no texto da política.

A Figura 4-4 ilustra a fase de análise com dois passos: (i) implementar o formalismo para casamento de padrões e (ii) analisar uma política de privacidade.

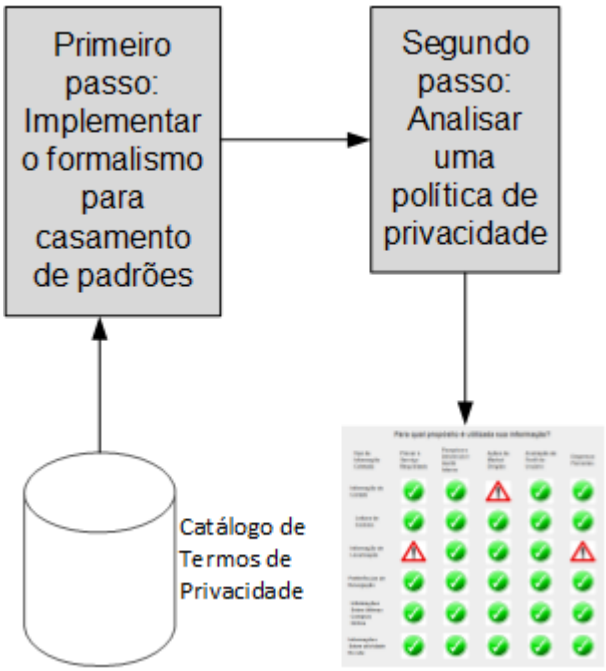

Figura 4-4: Visão geral da Fase de análise

4.3.2.1 Primeiro passo - Implementar o formalismo para casamento de padrões

Para analisar as políticas de privacidade foi definido um formalismo ao casamento de padrões. Para isso considere-se as definições a seguir.

Primeiramente, foram definidos os termos frequentes encontrados na fase de aprendizagem como conjunto A, sendo que:

$$A = conjunto\ de\ todos\ os\ termos\ chaves$$

O agrupamento dos termos frequentes em categorias é definido por A_i, onde:

$$A_i = \{t | t\ \in A\}$$

O conjunto de todas as categorias é definido por:

$$A = \bigcup_{i=1}^{n} A_i$$

Os conjuntos de categorias A_i são distintos entre eles, equivalendo a:

$$A_i \cap A_j = \emptyset \, para \, todo \, i, j \in \{1, n\} \, com \, i \neq j$$

Ou seja:

$$\bigcap_{i=1}^{n} A_i = \emptyset$$

Como descrito na primeira fase, os termos devem ser separados em duas classes: tipos de dados e tipos de propósitos. Dessa forma, o conjunto A é subdividido em dois conjuntos $Adado$ e $Aproposito$, logo o conjunto A é a união dos conjuntos $Adado$ e $Aproposito$, dado por:

$$A = Adado \cup Aproposito$$

Sendo que todo A_i pertence ao conjunto $Adado$ ou $Aproposito$, sendo representado por:

$$A_i \in (Adado) \, ou \, A_i \in (Aproposito)$$

Se $A_i \in Adado$, então se diz que a função $nome(A_i)$ denota uma informação clara ao usuário sobre qual tipo de dado das frases contém os termos dessa categoria. A função $nome(A_i)$ irá ser utilizada na construção do Rótulo de Privacidade.

Se $A_i \in Proposito$ então se diz que a função $nome(A_i)$ denota uma informação clara ao usuário do propósito das frases que contêm os termos desta categoria.

A função $nome(A_i)$ indica os termos que serão utilizados no Rótulo de Privacidade fornecido pelo especialista. Para $nome(A_i)$ quando $A_i \in Adado$, será utilizada uma linha da tabela do rótulo, já para $nome(A_i)$, quando $A_i \in Proposito$, será utilizada uma coluna da tabela do Rótulo.

Por exemplo: suponha-se que o conjunto A seja composto pelos termos {identidade, cpf, telefone, e-mail, localização, endereço, ofertas, propagandas}. Assim, pode se subdividir nos conjuntos $Adado$ e $Aproposito$, onde os termos "identidade, cpf, telefone, e-mail, localização e endereço" são tipos de dados e "ofertas e propagandas" são tipos de propósitos de utilização da coleta. Portanto,

$$A = \{\text{identidade, cpf, telefone, email, localização, endereço, ofertas, propagandas}\}$$
$$Adado = \{identidade, cpf, email, localização, endereço\}$$
$$Aproposito = \{ofertas, propagandas\}$$
$$A_1 = \{identidade, cpf, telefone, email\}$$
$$A_2 = \{localização, endereço\}$$
$$A_3 = \{ofertas, propagandas\}$$

De acordo com o exemplo acima, pode-se verificar que foram geradas três categorias “A_1, A_2 e A_3”, e, utilizando-se da função $nome(A_i)$, tem-se:

$$nome(A_1) \rightarrow Informação\ Pessoal\ e\ Contato$$
$$nome(A_2) \rightarrow Informação\ de\ Localização$$
$$nome(A_3) \rightarrow Ações\ de\ Marketing$$

A categoria A_1 recebeu o nome de "Informação pessoal e contato" por conter termos que identificam dados pessoais e de contato do usuário; a categoria A_2 recebeu o nome de "Informação de localização" por conter termos que identificam informações sobre localização do usuário; e a A_3 recebeu o nome de "Ações de marketing" por possuir termos que identificam o propósito de oferecer publicidade para o usuário. A Figura 4-5 ilustra um exemplo da utilização da função $nome(A_i)$ sobre os conjuntos de termos e categorias.

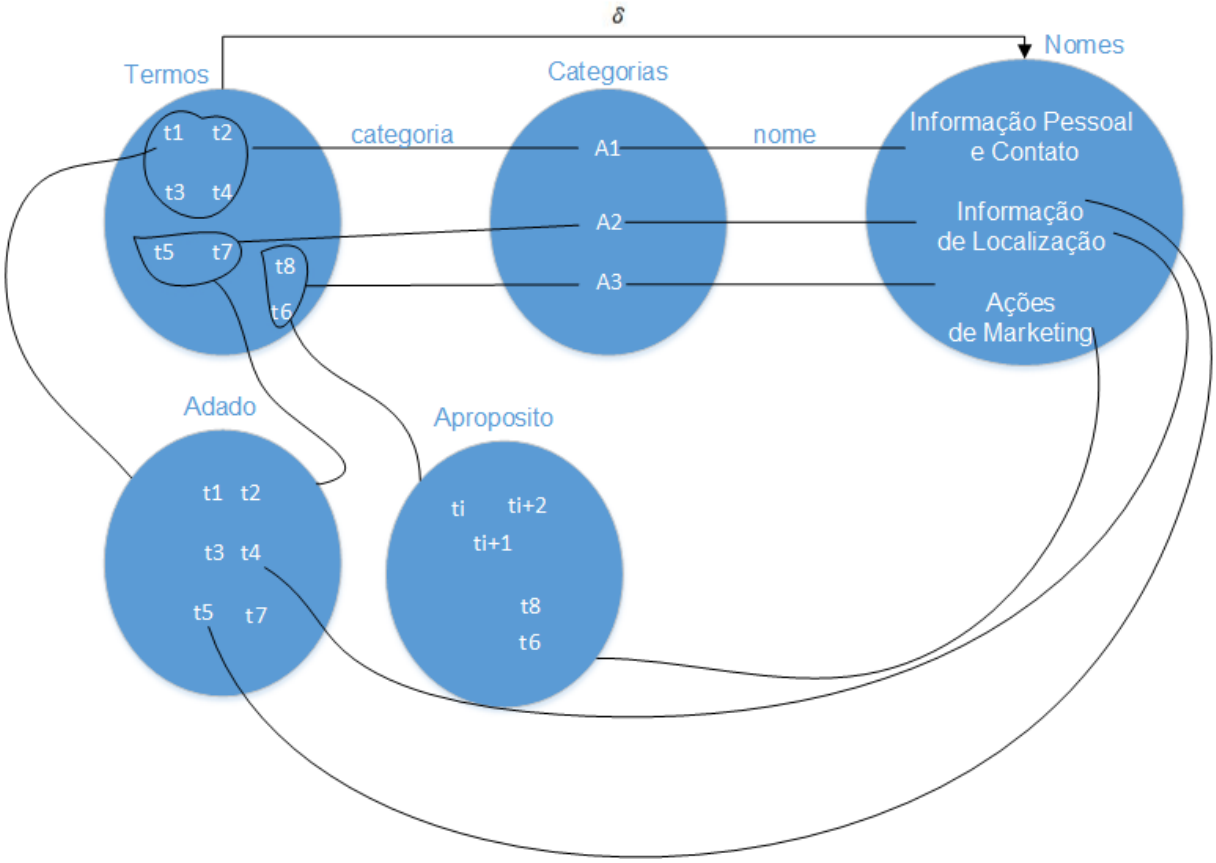

Figura 4-5: Ilustração da utilização das funções $nome(A_i)$, $categoria(t_i)$ e a composição de $\delta(t_i)$ sobre os conjuntos

Conforme ilustrado na Figura 4-5, além da função $\boldsymbol{nome(A_i)}$ existem outras duas funções, sendo as funções $categoria$ e δ. A função $categoria$, por meio dos termos t_i, define os conjuntos A_i, dados por:

$$categoria(t_i) \rightarrow Categorias$$

Já a função δ associa os termos t_i com o nome da categoria, sendo formada pela composição das funções $nome$ e $categoria$:

$$\delta(t_i) = nome \cdot categoria(t_i)$$

Para gerar o rótulo de privacidade de acordo com as definições descritas acima, é utilizado o produto cartesiano dos conjuntos A_i definido por:

$$RotPriv: Adado\ x\ Aproposito \rightarrow B\ onde\ B = \{N, C\}$$

O mapeamento $RotPriv$ é uma função que tem como domínio o produto cartesiano de $Adado$ com $Aproposito$ e tem como imagem o conjunto $B = \{N, C\}$.

4.3.2.2 Segundo passo - Analisar uma política de privacidade

Após definir o formalismo da categorização dos termos, é dado o formalismo para as ocorrências dos casamentos de padrões.

Seja $p = \{w_1, w_2\}$ onde w_1, w_2 são *tokens* obtidos do texto de entrada por um analisador léxico. Tais *tokens* são palavras da língua portuguesa separadas por sílabas ou sinais de pontuação.

O processo de casamento de padrões analisa os pares (w_i, w_j) para $i = \{1, \dots, n\}$ e $j = \{1, \dots, n\}$. Caso $w_i \in Termos$ e $w_j \in Termos$ então considere-se $categoria(w_i) = A_i$ e $categoria(w_j) = A_j$.

Se $A_i \in Adado$ e $A_j \in Aproposito$ ou $A_j \in Adado$ e $A_i \in Aproposito$, então tais pares identificam que o texto possui uma indicação a respeito da privacidade. Ou seja, um termo de dados pessoais casa com um termo de propósito de uso de tal dado. Assim, registra-se na tabela de Rótulo de Privacidade $RotPriv[i, j]$ com "C", informando que há aquele tipo de coleta e utilização do dado.

Caso a condição $A_i \in Adado$ e $A_j \in Aproposito$ ou $A_j \in Adado$ e $A_i \in Aproposito$ não for satisfeita, entende-se que no texto da política de privacidade não há informações sobre utilização das informações dos usuários. O exemplo a seguir ilustra o processo de casamento de padrões.

Utilizando-se das informações citadas anteriormente sobre o conjunto A, tem-se:

$$A = \{\text{identidade, cpf, telefone, email, localização, endereço, ofertas, propagandas}\}$$

$$Adado = \{identidade, cpf, telefone, email, localização, endereço\}$$

$$Aproposito = \{ofertas, propagandas\}$$

$$A_1 = \{identidade, cpf, telefone, email\}$$

$$A_2 = \{localização, endereço\}$$

$$A_3 = \{ofertas, propagandas\}$$

Seja o texto a ser analisado: “Podemos coletar seu telefone e e-mail para fornecer ofertas de nossos produtos”.

Sendo w o conjunto de *tokens*:

$$w = \{\text{podemos, coletar, seu, telefone, e, email, para, fornecer, ofertas, de, nossos, produtos}\}$$

Para aplicar o casamento de padrões, a entrada deve ser analisada em pares de *tokens* e para isto é necessário o produto cartesiano de w^2, resultando em:

$$\{(podemos, coletar), \dots, (podemos, produtos), \dots, (telefone, ofertas), \dots, (email, ofertas)\}$$

Então se inicia o processo do formalismo analisando os pares individualmente, onde o primeiro par é $(podemos, coletar)$, representado por $p = (w_1, w_2)$, sendo $w_1 = podemos$ e $w_2 = coletar$. A primeira análise é verificar se $w_i \; e \; w_j \in A$, lembrando que A representa o conjunto de todos os termos. Neste caso:

$$p = (w_1, w_2) \in A$$

onde

$$p = (podemos, coletar)$$
$$\in \{\text{identidade, cpf, telefone, email, localização, endereço, ofertas, propagandas}\}$$

Para $w_{i=1} \; e \; w_{j=2}$ não há casamento com nenhum elemento do conjunto A, então$p = (w_{i=1}, w_{j=2}) \cap A = \emptyset$. A análise do par (w_1, w_2)é finalizada, pois não há nenhum termo de p em A. Cada par do produto cartesiano w^2 é analisado dessa forma. Tomando outro exemplo para p, onde $p = (w_6, w_9)$, sendo $w_6 = email$ e $w_9 = ofertas$, tem-se

$$p = (w_6, w_9) \in A$$

onde

$$p = (email, ofertas)$$
$$\in \{\text{identidade, cpf, telefone, email, localização, endereço, ofertas, propagandas}\}$$

Já para $w_{i=6}$ e $w_{j=9}$ há casamento com elementos do conjunto A, então $p = (w_{i=1}, w_{j=2}) \cap A = \{email, ofertas\}$. Como há elementos de p em A, deve ser analisado em qual conjunto ($Adado\ ou\ Aproposito$) eles são pertencentes, sendo os conjuntos:

$$Adado = \{identidade, cpf, telefone, email, localização, endereço\}$$
$$Aproposito = \{ofertas, propagandas\}$$
$$A_1 = \{identidade, cpf, telefone, email\}$$
$$A_2 = \{localização, endereço\}$$
$$A_3 = \{ofertas, propagandas\}$$

Recordando que $categoria(w_i) = A_i$ e $categoria(w_j) = A_j$, onde $w_{i=6} = email$ e $w_{j=9} = ofertas$, utilizando a função $categoria(w_i)$ para $categoria(w_6) = A_1$, pois o termo *e-mail* é um elemento da categoria A_1 e para $categoria(w_9) = A_3$, pois o termo *ofertas* é um elemento de A_3.

Nesse momento já foi determinado a qual ou a quais categorias os termos pertencem. Agora é necessário verificar o tipo de informação fornecida pelas categorias. Para isso analisa-se:

$$A_i \in Adado \text{ e } A_j \in Aproposito \text{ ou } A_j \in Adado \text{ e } A_i \in Aproposito$$

Deste modo temos que $A_1 \in Adado$ e $A_3 \in Aproposito$, casando corretamente um tipo de informação do usuário com um propósito descrito. Com essas informações deve ser aplicada a função $nome(A_i)$ para ambas as categorias, resultando em:

$$nome(A_1) \rightarrow Informação\ Pessoal\ e\ Contato$$
$$nome(A_3) \rightarrow Ações\ de\ Marketing$$

Dessa forma é possível mapear o resultado no Rótulo de Privacidade com a função $RotPriv[1,3] \rightarrow$ ✅.

O processo de formalismo deve ser aplicado para todos os pares do produto cartesiano de w^2, sendo que para cada ocorrência de coleta de um dado do usuário

($Adado$) associado com alguma finalidade ($Aproposito$) o Rótulo de Privacidade deve ser preenchido com e para os casos em que não houver ocorrência de casamento de padrões entre dado e finalidade, o Rótulo deve ser preenchido com ⚠.

Para identificar no Rótulo se há algum tipo de coleta e utilização de dados dos usuários, foram definidos dois símbolos pertencentes ao conjunto B representando ✅ para “está sendo coletado aquele tipo de informação e para aquele propósito” e ⚠ para “não está sendo coletado ou não foi encontrado no texto da política aquele tipo de informação e para aquele propósito”. A Figura 4-6 ilustra o protótipo da legenda elaborada.

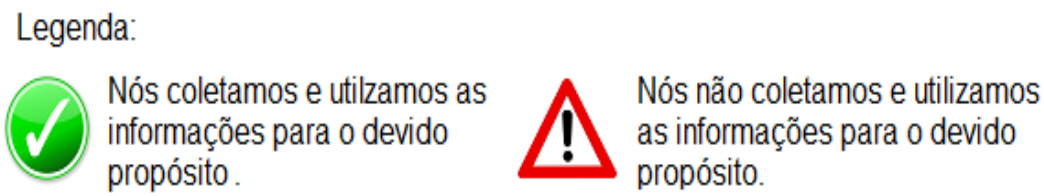

Figura 4-6: Legenda para identificação da coleta e utilização dos dados dos usuários

A Figura 4-7 ilustra a forma como o Rótulo de Privacidade é preenchido com as legendas elaboradas.

De acordo com a ilustração da Figura 4-7, pode-se verificar a presença da legenda no Rótulo de Privacidade. A ilustração representa que estão sendo coletadas informações pessoais e contatos para ações de marketing e informações de localização para prover o serviço solicitado e o compartilhamento com empresas terceiras.

Figura 4-7: Exemplo de Rótulo com a legenda

4.4 Considerações finais

A metodologia descrita neste capítulo permite encontrar palavras-chaves de privacidade em um conjunto de textos com o auxílio de um especialista. Com essas palavras-chaves são criadas categorias de privacidade, definindo os tipos de informações dos usuários que podem ser coletadas e também o propósito da utilização destas informações.

Com a utilização do formalismo proposto, pode-se analisar uma política de privacidade e o processo busca encontrar padrões das palavras-chaves no texto, caso houver ocorrências de padrões na análise da política. Essas ocorrências são

mapeadas em uma tabela para apresentação ao usuário, servindo de suporte para conscientização de quais dados são coletados pelo serviço on-line.

Capítulo 5

PPMARK – UM PROTÓTIPO DE FERRAMENTA PARA GERAÇÃO DE RÓTULO DE PRIVACIDADE

Com o objetivo de validar a técnica apresentada para geração de Rótulo de Privacidade deste trabalho, foi desenvolvido um protótipo de uma aplicação utilizando os passos da Fase de análise.

5.1 Considerações iniciais

Para empregar o formalismo de casamento de padrões descrito no Capítulo 4, foi desenvolvido um protótipo de uma aplicação para analisar as políticas dos serviços on-line, tornando possível a geração do Rótulo de Privacidade automaticamente. A Figura 5-1 ilustra a visão geral do desenvolvimento do protótipo da aplicação.

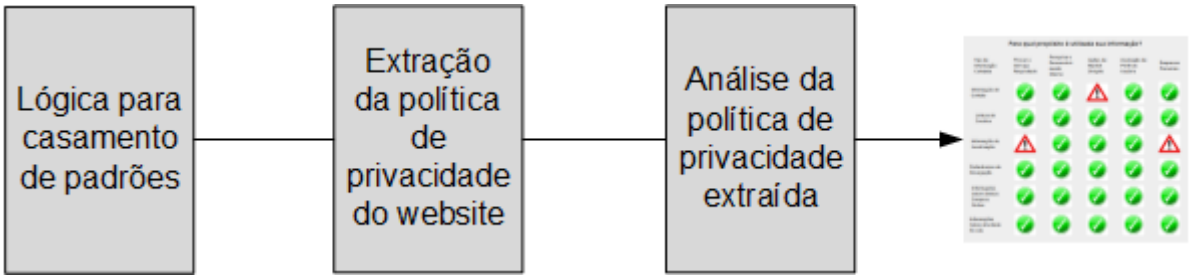

Figura 5-1: Visão geral da aplicação PPMark

5.2 Implementação do formalismo para casamento de padrões

Primeiramente, deve-se implementar o casamento de padrões para análise das políticas dos serviços on-line. No desenvolvimento desse passo foi utilizado o algoritmo *Rabin-Karp*[16] para localizar vários conjuntos de padrões em um determinado texto. O Código-fonte 5-1 mostra a forma como foi implementado no protótipo o processo de casamento de padrões.

```
1  ...
2  for (String celulaTabela :  celulaTabelaOrdenada) {
3      int qtdPositivos = 0;
4      int qtdNegativos = 0;
5      int contadorParagrafo = 1;
6      for (String paragrafo : conteudoArquivoPolitica) {
7          String resultadoCelula = "";
8          if(paragrafoAceito(paragrafo)){//paragrafo não vazio
9              int qtdDePadroesYEncontrado = Tabela.analisaYdaCelula(
                   celulaTabela, paragrafo);
10             if(qtdDePadroesYEncontrado > 0){
11             int qtdDePadroesXEncontrado = Tabela.analisaXdaCelula(
                   celulaTabela, paragrafo);
12                 if (qtdDePadroesXEncontrado > 0) {
13                     qtdPositivos++;
14                 }else{
15                     qtdNegativos++;}
16             }else{
17                 qtdNegativos++;}
18         contadorParagrafo++;
19         }
20     resultadoCelula = "";
21     }
22 }
23 ...
```

Código-fonte 5-1: Implementação do casamento de padrões

A lógica do funcionamento do casamento de padrões se dá da seguinte maneira: inicialmente com interações em cada célula da tabela (linha 2), sendo que

[16]Implementação disponível em: https://github.com/sarveshsaran/RabinKarp

cada célula representa uma categoria de privacidade. Posteriormente, o conteúdo do arquivo é percorrido, analisando cada parágrafo do texto (linha 6). O algoritmo verifica se o parágrafo não é vazio (linha 8). Caso não seja, o algoritmo irá procurar no parágrafo algum termo relacionado a tipo de dados (linha 9).

Após analisar o parágrafo em busca desses termos é verificado se houve alguma ocorrência de casamento de padrão (linha 10). Caso seja verdadeira a condição, então se inicia a busca no mesmo parágrafo, procurando algum termo sobre utilização dos dados (linha 11). Também é verificado se foi encontrado algum termo daquela célula (linha 12). Caso seja encontrado, é contabilizado que no parágrafo em análise foram encontrados tanto termos de tipo de dados quanto utilização dos dados (linha 13). Assim, a devida célula da tabela é preenchida com a letra “C”, que representa que estão sendo coletados e utilizados os dados do usuário.

Caso não ocorra nenhum casamento de padrão dos termos de tipo de dados, o algoritmo já entende que não haverá ocorrência naquela célula, então ele já contabiliza que não houve casamento de padrão (linhas 10 e 17). A devida célula é preenchida com a letra “N”, que representa que não estão sendo coletados e utilizados os dados do usuário ou que não foram encontrados os termos no texto.

Desse modo, o algoritmo passa para o próximo parágrafo (linha 18), analisando as mesmas células. Essas interações percorrem o texto todo. Encerrando a análise em todos os parágrafos, o algoritmo recebe outra célula para análise (linha 2), realizando desta forma, sucessivamente, a comparação para todas as células em todos os parágrafos do texto.

A utilização da implementação do algoritmo *Rabin-Karp* encontra-se nos métodos *analisa Y da Célula* e *analisa X da Célula*, mostrados nas linhas 9 e 11 do Código-fonte 5-1. Os métodos fazem chamadas do método *acha Padrão* que, por sua vez, invoca o algoritmo *Rabin-Karp*. O método *acha Padrão* recebe um parágrafo e um padrão a ser comparado.

Para a execução do algoritmo *Rabin-Karp* são necessários dois parâmetros passados pelo método *acha Padrão*. O primeiro deve ser uma entrada para ser comparada e o segundo um padrão já definido, conforme é ilustrado na linha 4 do Código-fonte 5-2.

```
1 ...
2 public static int achaPadrao(String entry, String pattern) {
3     int retorno = 0;
4     RabinKarp RK = new RabinKarp(entry, pattern);
5     return RK.isSubstring();
6 }
7 ...
```

Código-fonte 5-2: Chamada da função de casamento de padrões com Rabin-Karp

Ao ser chamado o método *isSubstring*, o algoritmo processa a entrada denominada *entry* e o padrão *pattern*. Esse processamento retorna como resposta um valor maior ou igual a 0, caso o padrão seja encontrado, e o valor -1, caso o padrão não seja encontrado. Com isso é possível denominar se o padrão foi encontrado ou não, conforme é observado no Código-fonte 5-1, linhas 10 e 12.

Após implementar a lógica para casamento de padrões, é possível executar o segundo passo responsável por extrair a política de privacidade do website, fazer o processamento no texto e executar a implementação descrita, gerando a tabela com os respectivos resultados dos casamento de padrões.

5.3 Extração da política de privacidade do website

Após implementada a lógica de casamento de padrões descrita anteriormente, pode-se iniciar o processo de análise de uma política de privacidade de um website. Para tanto, se faz necessário, primeiramente, encontrar o texto da política. Dessa forma, é preciso realizar uma procura dentro do código-fonte HTML do website e achar o link que o represente e consequentemente o texto da política de privacidade. Com esse intuito foi desenvolvido um método que percorre todo o HTML do website procurando pelo link da política e utilizando a biblioteca *Jsoup*[17]. Esta biblioteca foi desenvolvida especificamente para manipular páginas em HTML. O Código-fonte 5-3 ilustra como é feita a busca pelo link da política de privacidade.

[17]Disponível em: https://jsoup.org/

```
1  ...
2  private String retornaLink(Elements links){
3      String linkEncontrado = "";
4      for (Element link : links) {
5          if (link.text().contains("privacidade") || link.text().contains("
               Privacidade") || link.text().contains("política") || link.text()
               .contains("politica")
6               || link.text().contains("Política") || link.text().contains("
                   Politica")){
7          linkEncontrado = link.attr("abs:href");
8          }
9      }
10     return linkEncontrado;
11 }
12 ...
```

Código-fonte 5-3: Localização do link da política de privacidade

Conforme pode ser visto com início na linha 4 do Código-fonte 5-3, o algoritmo, com o auxílio dos métodos da biblioteca *Jsoup,* percorre todos os links do serviço on-line buscando os termos *"privacidade, Privacidade, política, politica, Política e Politica"* para a língua portuguesa. Quando é encontrado algum dos termos, significa que a política de privacidade se encontra naquele link e o método *link.attr ("abs:href")* presente na linha 7 captura o link.

Depois de encontrar o link, utilizando-se do método *parse* demonstrado no Código-fonte 5-4, faz-se a conversão do arquivo HTML para arquivo de texto. Dessa forma é possível se obter um texto limpo e sem as *tags* do HTML. A biblioteca *Jsoup* auxilia todo esse processo, realizando-o automaticamente, por meio de seus métodos.

```
1  ...
2  public void parse(Reader in) throws IOException {
3      s = new StringBuffer();
4      ParserDelegator delegator = new ParserDelegator();
5      delegator.parse(in, this, Boolean.TRUE);
6  }
7  ...
```

Código-fonte 5-4: Método para conversão de HTML para texto

Ao se obter o texto da política de privacidade é feito um pré-processamento. Nele, todos os caracteres especiais, acentuações e *stopwords* são removidos, além de todas as palavras serem convertidas para letras minúsculas. A motivação atrás dessa conversão é o fato de, após realizada, facilitar o casamento de padrões pelo algoritmo citado anteriormente. O Código-fonte 5-5 demonstra o trecho do código que faz o pré-processamento do texto.

```
1  ...
2  public String clean(String text){
3      //processo 1: transformar todos os caracteres em minusculo
4      text = text.toLowerCase();
5      //processo 2: eliminar caracteres fora de A-Z
6      text = eliminateDisallowedCharacters(text);
7      //processo 3: eliminar stopwords
8      text = removeStopwords(text);
9      return text;
10 }
11 ...
```

Código-fonte 5-5: Pré-processamento do texto para análise

De acordo com o Código-fonte 5-5, o conteúdo do texto sofre três processos, sendo que o primeiro converte todas as letras maiúsculas para letras minúsculas (linha 4), o segundo remove todos os caracteres especiais (linha 6) e por fim remove todas as *stopwords* (linha 8).

5.4 Analisar uma política de privacidade

Com o texto preparado e pré-processado, o protótipo irá executar o casamento de padrões implementado conforme descrito na seção 5.2, verificando se cada parágrafo tem as palavras-chaves. Caso ocorram os padrões especificados, a tabela será gerada com as informações extraídas. A Figura 5-2 ilustra a interface do protótipo PPMark.

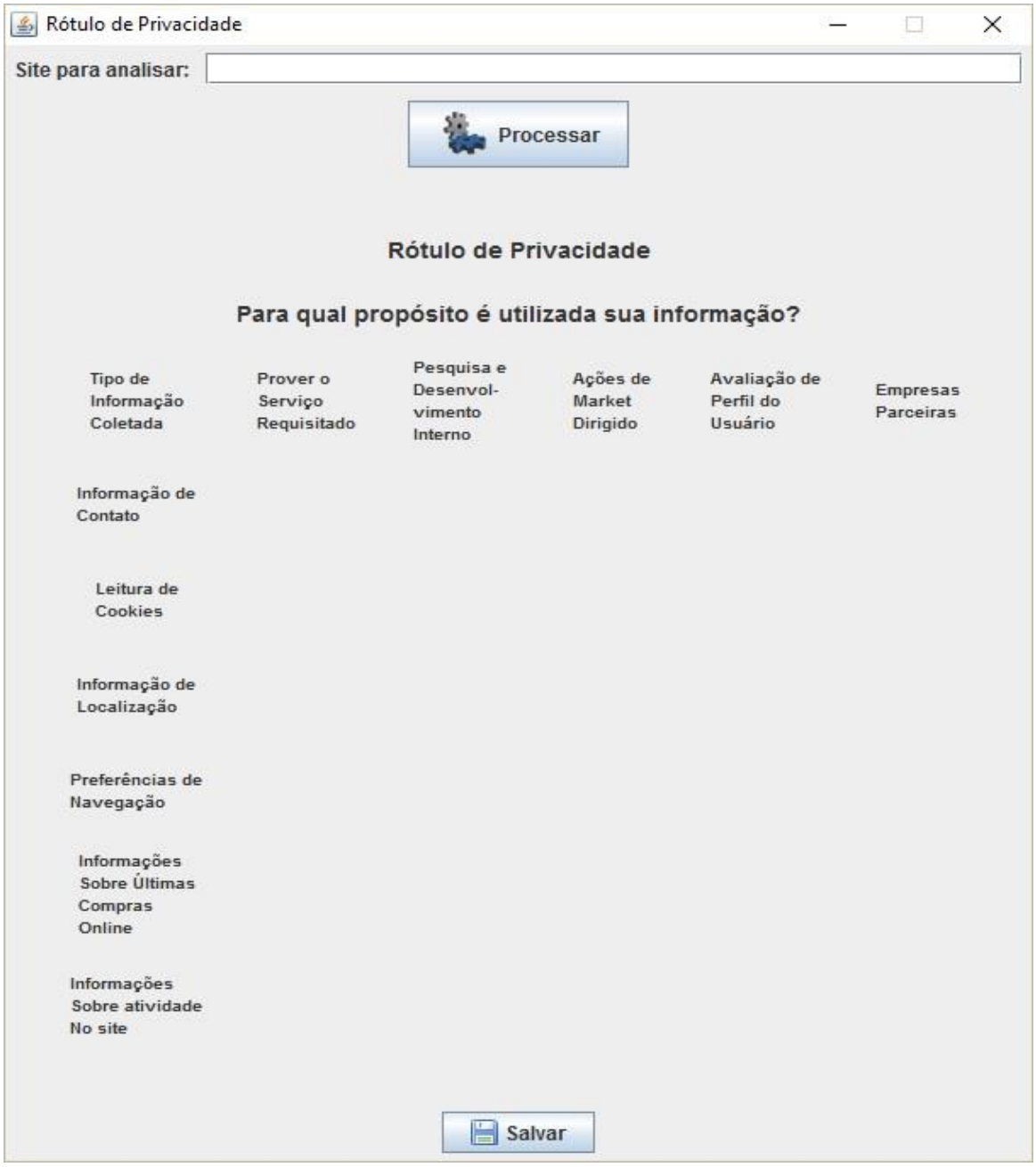

Figura 5-2: Interface do protótipo da PPMark

Conforme pode ser observado na Figura 5-2, a interface contém um campo para digitar o endereço do website e um botão para iniciar o processamento da política. Quando pressionado o botão “Processar”, o protótipo inicia o processamento citado anteriormente, bem como a execução da lógica de casamento de padrões. É possível, assim, o preenchimento do Rótulo, conforme exemplificado na Figura 5-3.

Figura 5-3: Interface do protótipo da PPMark com possíveis resultados

5.5 Considerações finais

Neste capítulo foram abordados breves conceitos sobre o desenvolvimento do protótipo da aplicação PPMark. O desenvolvimento do protótipo foi utilizado para avaliar a técnica apresentada neste trabalho. A ferramenta é capaz de acessar um serviço on-line e encontrar a política de privacidade, desde que haja uma política disponível no código-fonte HTML. O processo apresentará as informações sobre coleta e utilização de dados citadas na política do serviço, conforme demonstrado pelas Figura 5-2 e Figura 5-3.

Capítulo 6

AVALIAÇÃO

Neste capítulo serão abordados conceitos sobre a condução da avaliação da metodologia apresentada neste trabalho, tais como o método de avaliação e os resultados obtidos na geração de Rótulo de Privacidade dos serviços on-line.

6.1 Considerações iniciais

Kelley et al. (2010) propuseram um experimento para avaliar os formatos em que as políticas de privacidade são apresentadas aos usuários. Os autores verificaram que o rótulo de privacidade pode facilitar o entendimento sobre os termos das políticas de privacidade dos serviços virtuais.

Como o objetivo deste trabalho foi adaptar o Rótulo de Privacidade elaborado por Kelley et al. (2009), fazendo seu preenchimento automaticamente com a extração de informações dos textos das políticas de privacidade que são escritos em linguagem natural, o foco da avalição é medir a precisão que as informações sobre coleta de dados nas políticas são recuperadas.

Foram utilizadas três métricas diferentes para avaliar a recuperação das informações (LANDGREBE et al., 2006; LIU e ÖZSU, 2009): (i) precisão, que é baseada na noção de itens classificados corretamente; (ii) *recall*, cálculo da porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas e (iii) *f-measure*, que é utilizada para avaliar a exatidão de um classificador, fazendo a ponderação entre precisão e *recall* (RONCERO, 2010).

6.2 Metodologia da avaliação

Para realizar essa avaliação foi selecionada uma amostra de políticas de privacidade. A avaliação foi estruturada em cinco fases, que são ilustradas na Figura 6-1.

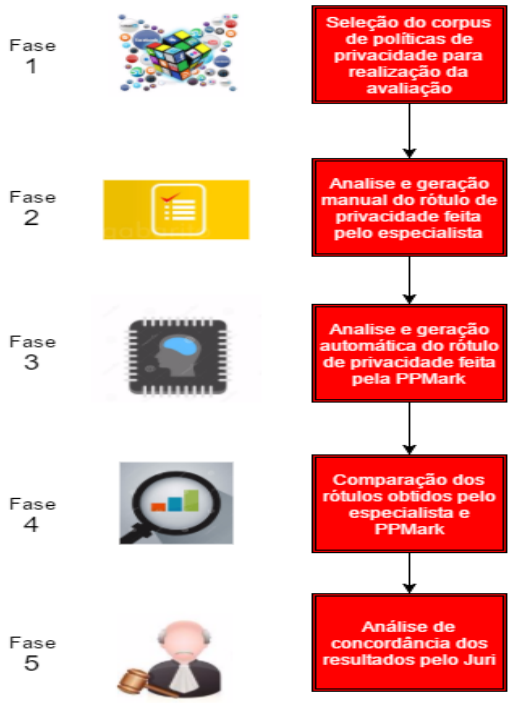

Figura 6-1: Fases da avaliação da metodologia

A primeira fase foi a seleção do corpus de políticas de privacidade para realizar a avaliação. Esse corpus foi constituído por uma amostragem aleatória simples sobre o conjunto de políticas selecionados no Capítulo 4 - . A escolha da amostragem aleatória simples sobre o conjunto de políticas previamente selecionadas se deu pelo fato de não se conhecer a quantidade total de serviços on-line disponíveis. Entretanto, foram selecionadas as 60 mais acessadas no Brasil, que certamente podem representar melhor as demais, devido ao número de usuários que já as utilizam e às funcionalidades aplicadas a elas, em decorrência do número de acessos e das necessidades impostas pelos usuários (LOBATO e ZORZO, 2007b).

A segunda fase foi a análise e geração manual do Rótulo de Privacidade, executada pelo especialista, na qual ele teve de analisar a amostra de políticas de

privacidade, localizando nos textos informações sobre coleta e utilização dos dados dos usuários. Com isso os Rótulos de Privacidade foram gerados manualmente, sendo assinalada cada ocorrência das categorias de privacidade citadas nas políticas da amostra.

A terceira fase foi a análise e geração automática do Rótulo de Privacidade, executada pela PPMark, que, após o especialista analisar a amostra, teve de ser analisada pelo protótipo da aplicação PPMark, sendo os Rótulos de Privacidade gerados de forma automática.

A quarta fase foi a comparação dos rótulos gerados pela execução da análise do especialista com os gerados pelo protótipo da aplicação. Os resultados foram confrontados e as métricas de precisão, *recall* e *f-measure* aplicadas nos resultados obtidos.

Por fim, a quinta fase foi uma avaliação feita por um júri sobre os resultados obtidos. Esta avaliação foi baseada na porcentagem de concordância absoluta (*percentage of absolute agreement*) (MATOS, 2014). Como as políticas de privacidade foram analisadas pelo especialista e a análise é passível de interpretação, o jurado fez a mesma análise para classificar se está ou não de acordo com a interpretação do especialista. Cada fase será detalhada nas subseções seguintes.

6.2.1 Primeira fase - amostragem

Inicialmente, definiu-se o tamanho da amostra sobre o conjunto de textos selecionados no Capítulo 4 - utilizando-se o cálculo de amostragem. Considerando-se o tamanho selecionado da população de 60 políticas, sendo que as políticas de privacidade são homogêneas, pois retratam informações sobre privacidade das informações dos usuários, com nível de confiança de 90% e uma margem de erro de 5%, obteve-se o valor de dez políticas de privacidade para a realização dos testes. Para escolher as dez políticas no corpus empregou-se amostragem aleatória simples, com utilização da tabela de números aleatórios.

As políticas selecionadas contemplaram os segmentos de e-commerce, noticiários, serviços de e-mail, bancários e *streaming*. A Tabela 6-1 apresenta esses estratos. Para preservar o anonimato das empresas foram omitidos os nomes dos serviços on-line.

Tabela 6-1: Segmentos dos serviços on-line utilizados para testes

Segmento	Quantidade	Arquivos (.txt)
E-commerce	5	política 01, política 03, política 08, política 09, política 10
Serviço bancário	1	política 06
Noticiários	2	política 04, política 07
Serviço de streaming	1	política 02
Serviço de e-mail	1	política 05

6.2.2 Segunda fase - Análise e geração manual do Rótulo de Privacidade dos serviços selecionados

Após determinar a amostra para análise, as políticas de privacidade foram analisadas manualmente. A análise visou verificar se as categorias de privacidade sobre coleta e utilização de dados estavam descritas nos textos das políticas.

Para isso, o especialista gerou um Rótulo de Privacidade para cada política, anotando as categorias existentes. O número de ocorrências geral das categorias é apresentado na Tabela 6-2. Os Rótulos de Privacidade gerados pelo especialista podem ser acessados virtualmente[18].

Tabela 6-2: Número de categorias de privacidade relevantes encontradas nos textos pelo especialista

Política	Quantidade de categorias relevantes encontradas
Política1	27
Política2	18
Política3	17
Política4	6
Política5	7
Política6	6
Política7	7
Política8	17

[18] Os rótulos gerados pelo especialista e pela aplicação estão disponíveis para acesso em: https://www.dropbox.com/sh/jyamh3al7tcy7qj/AADUFXZ3shvQ2DlDYoX6XQTza?dl=0

Política9	13
Política10	15
Total	133

De acordo com a Tabela 6-2, foram contabilizadas 133 ocorrências das categorias de coleta e utilização de dados na amostra de políticas de privacidade pelo especialista.

Após a geração dos rótulos de cada política e a contabilização das categorias, foram gerados os rótulos das mesmas políticas de forma automatizada pelo protótipo da aplicação PPMark, que é descrito na subseção seguinte.

6.2.3 Terceira fase - Análise e geração automática do Rótulo de Privacidade dos serviços selecionados

Nessa etapa foi utilizado o protótipo da aplicação PPMark para gerar os rótulos automaticamente, registrando as categorias que o protótipo conseguiu extrair. A Tabela 6-3 apresenta o número geral de ocorrências de categorias de privacidade recuperadas pela PPMark.

Com a utilização do protótipo da aplicação PPMark, conforme pode ser visualizado na Tabela 6-3, foram contabilizadas 113 ocorrências relevantes das categorias de coleta e utilização de dados na amostra de políticas de privacidade.

Após a análise automática e a geração dos rótulos, foram comparados os resultados da geração dos rótulos feita pelo especialista e os gerados pela aplicação. A comparação é descrita na subseção seguinte.

Tabela 6-3: Número de categorias de privacidade relevantes encontradas nos textos automaticamente

Política	Quantidade de categorias encontradas
Política1	19
Política2	15
Política3	15
Política4	6
Política5	7
Política6	5

Política7	7
Política8	12
Política9	12
Política10	15
Total	113

6.2.4 Quarta fase - Comparação entre os rótulos gerados pela execução da análise pelo especialista e os gerados pelo protótipo da aplicação

Após a geração manual e automática dos rótulos, foram comparados os resultados com o objetivo de avaliar a precisão da recuperação das categorias de privacidade.

Para avaliar os resultados entre as comparações, foram utilizadas três métricas de avaliação da recuperação de informação, sendo: (i) precisão, que é baseada na noção de itens classificados corretamente; (ii) *recall*, cálculo da porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas e (iii) *f-measure*, que é utilizada para avaliar a exatidão de um classificador, fazendo a ponderação entre precisão e *recall* (RONCERO, 2010; SEBASTIANI, 2002).

Para ser possível calcular as três métricas que serão apresentadas na Seção 6.2.5, são necessárias algumas definições para efetuar os cálculos, nos quais: categorias relevantes recuperadas manualmente, nomeadas como *Ground Truth* - **GT,** representam as categorias que foram localizadas pelo especialista; características relevantes recuperadas, nomeadas como *True Positive* – **TP,** representam as categorias recuperadas verdadeiras, ou seja, existentes nos textos e recuperadas pelo protótipo; características relevantes não recuperadas, nomeadas como *False Negative* - **FN,** representam as características verdadeiras não recuperadas, ou seja, existentes nos textos, porém não recuperadas pelo protótipo e características irrelevantes recuperadas, nomeadas como *False Positive* - **FP,** representam as características não verdadeiras recuperadas, ou seja, não existentes nos textos, porém encontradas pelo protótipo.

De acordo com as definições, foram contabilizadas as ocorrências de cada item descrito, sendo que a Tabela 6-4 apresenta as categorias de privacidade

recuperadas manualmente (GT), as categorias de privacidade relevantes recuperadas pelo protótipo (TP), as categorias de privacidade relevantes não recuperadas (FN) e por fim as irrelevantes recuperadas (FP).

Tabela 6-4: Relação entre as categorias recuperadas manualmente e automaticamente

Política	GT	TP	FN	FP
Política1	27	19	8	1
Política2	18	15	3	3
Política3	17	15	2	0
Política4	6	6	0	4
Política5	7	7	0	4
Política6	6	5	1	1
Política7	7	7	0	5
Política8	17	12	5	0
Política9	13	12	1	0
Política10	15	15	0	8
Total	133	113	20	26

Após analisar cada política de privacidade, gerar os rótulos manual e automaticamente, contabilizar as ocorrências das categorias recuperadas verdadeiras ou não verdadeiras pelo protótipo da aplicação PPMark, foram calculadas as métricas de precisão, *recall* e *f-measure* com as informações apresentadas na Tabela 6-4, que são detalhadas na Seção 6.3.

6.2.5 Quinta fase - Análise de concordância dos resultados pelo Júri (Método do Júri)

A quinta fase teve como objetivo avaliar a interpretação do especialista nas análises textuais das políticas de privacidade. Para isso utilizou-se o Método do Júri, que compõe um jurado com dois juízes e estes analisam os resultados individualmente.

Após a análise dos juízes foi verificado se houve “concordância entre juízes” (MATOS, 2014). Para calcular a concordância foi empregada a porcentagem de

concordância absoluta (*percentage of absolute agreement*), que, segundo Matos (2014), é a técnica mais simples utilizada. A porcentagem de concordância absoluta consiste unicamente em calcular o número de vezes em que os avaliadores concordam e dividir pelo número total de avaliações, podendo o resultado variar entre 0 *e* 100% (FONSECA, SILVA e SILVA, 2007; MATOS, 2014).

A metodologia do emprego do cálculo de concordância foi posta em prática da seguinte maneira: (i) foram selecionados dois especialistas em privacidade para compor o jurado; (ii) foram entregues aos jurados os textos das políticas utilizadas nos testes de precisão; (iii) foram entregues os rótulos gerados manualmente pelo especialista; (iii) foram especificadas duas categorias para a classificação, sendo Concordo e Não concordo; (iv) foi elaborada uma planilha contendo os nomes dos arquivos das políticas e colunas para avaliação dos jurados com as categorias de classificação; (v) foi informado aos jurados que eles deveriam ler as políticas de privacidade e verificar se o rótulo de cada uma foi preenchida com as devidas informações de coleta e utilização de dados descritas nos textos, assim classificando como Concordo ou Não concordo e (vi) foi entregue aos jurados o “Termo de Consentimento Livre e Esclarecido” para registrar a sua participação nas avaliações.

6.3 Resultados

De acordo com os dados da Tabela 6-4, pôde-se verificar que houve uma aproximação das categorias relevantes (GT) para as categorias recuperadas pela técnica (TP). Essa aproximação pode ser melhor visualizada na Figura 6-2.

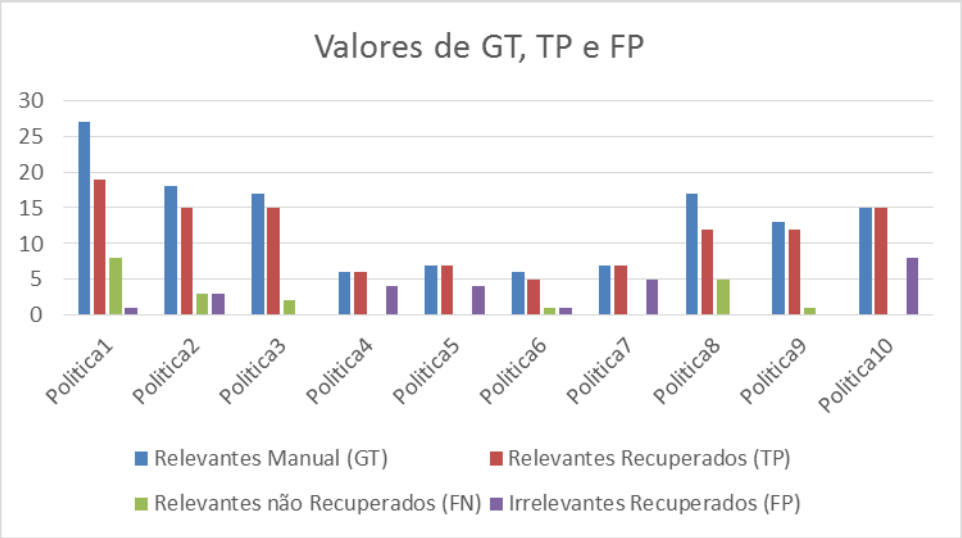

Figura 6-2: Aproximação das categorias relevantes (GT) e categorias recuperadas (TP)

Conforme pode ser visto na Figura 6-2, ao ser aplicada a técnica desenvolvida neste trabalho conseguiu-se recuperar categorias relevantes com uma aproximação média de 85% no geral. Essas categorias relevantes são marcadas na Figura 6-2 na cor vermelha (TP). Já as categorias encontradas por um especialista são marcadas pela cor azul (GT).

Com os resultados mostrados na Tabela 6-4 pôde-se calcular as métricas de avaliação da recuperação das categorias de privacidade, demonstradas pelas Equações 2, 3 e 4.

$$Precisao = \frac{TP}{GT + FP} \tag{2}$$

$$Recall = \frac{TP}{GT + FN} \tag{3}$$

$$F - Measure = 2 * \frac{Precisao * Recall}{Precisao + Recall} \tag{4}$$

onde GT foram as ocorrências encontradas pelo especialista, TP os *True Positives*, FP os *Falses Positives* e FN os *False Negatives*. Desta forma tem-se:

$$Precisao = \frac{113}{133 + 26} = 0{,}7106$$

$$Recall = \frac{113}{133 + 20} = 0{,}7385$$

$$F - Measure = 2 * \frac{0{,}7106 \ * \ 0{,}7385}{0{,}7106 \ + \ 0{,}7385} = 0{,}7244$$

Como o objetivo deste trabalho foi recuperar as categorias de privacidade citadas nos textos utilizando-se de palavras-chaves, focou-se em avaliar o quanto é precisa a técnica desenvolvida no sentido de recuperar tais informações. Para isso utilizou-se referências de valores para precisão abordados por Pérez-Castillo et al. (2011), apresentados na Tabela 6-5.

Tabela 6-5: Métricas para avaliação de precisão adaptada de Pérez-Castillo et al. (2011)

Se	P	=	0.47			=	Precisão muito baixa
Se	0.47	<	P	=	0.56	=	Precisão baixa
Se	0.56	<	P	=	0.63	=	Precisão média
Se	0.63	<	P	=	0.72	=	Precisão alta
se	P	>	0.72			=	Precisão muito alta

Considerando a média de ocorrências de categorias (TP) recuperadas no corpus de teste, o valor da precisão dada pela Equação 2 foi de $0{,}7106$ (71%), podendo-se afirmar, assim, que a técnica desenvolvida pode ser usada com um certo grau de confiança, uma vez que, de acordo com Pérez-Castillo et al. (2011), o valor de precisão é considerado alto.

O cálculo do *recall* é a relação das informações relevantes encontradas sobre todas as informações relevantes descritas nos textos. O valor obtido pela técnica para o cálculo de *recall* foi de $0{,}7385$ (74%). Ou seja, de 100% de todas as categorias de privacidade descritas no corpus de teste, foram recuperadas 74%.

Por fim tem-se a medida *F-measure*, ou medida F, que é utilizada para avaliar a exatidão de um classificador, fazendo a ponderação entre precisão e *recall*. O valor da medida F foi de $0{,}7244$, aproximadamente 0,72 (72%). O resultado da

medida F é um indicativo de que, quanto mais próximo de 100%, melhor é o desempenho (GUELPELI, BERNARDINI e GARCIA, 2008; RONCERO, 2010; SEBASTIANI, 2002). A Figura 6-3 ilustra a relação entre as três métricas de avaliação.

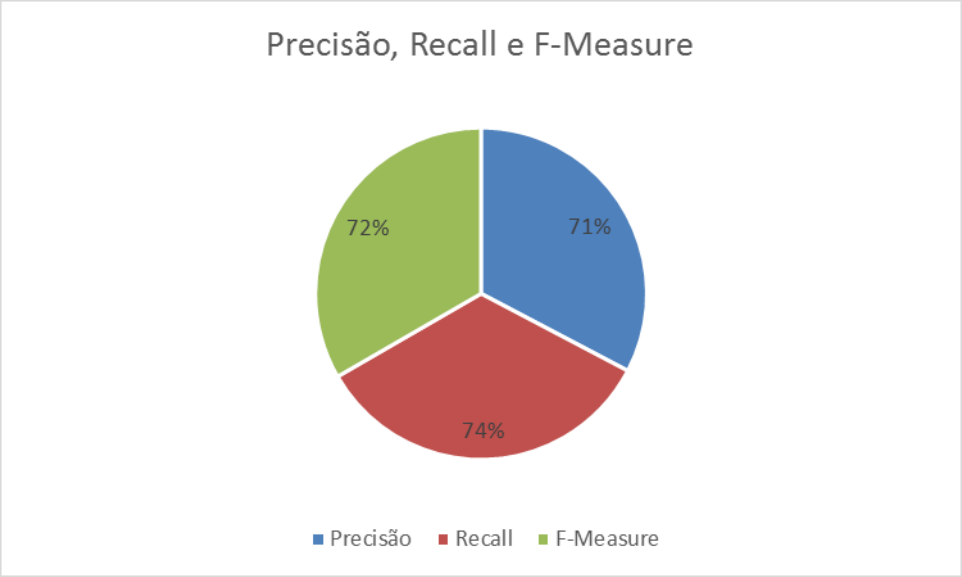

Figura 6-3: Relação entre Precisão, *Recall* e *F-Measure*

6.3.1 Avaliação do Júri

Para poder avaliar os rótulos gerados automaticamente, primeiro foi preciso analisar as políticas de testes manualmente, gerando, assim, os Rótulos de Privacidade pela interpretação do especialista. Porém, as políticas de privacidade podem ser interpretadas de várias maneiras. Isso pode ocorrer pelo fato de as políticas serem escritas em linguagem natural.

Para verificar a interpretação do especialista sobre os textos das políticas de privacidade, foram convidados dois juízes para analisar os rótulos gerados manualmente. Os juízes são mestres em Ciência da Computação e especializados na área de privacidade.

O jurado ficou incumbido de analisar tais rótulos e assim verificar se há concordância entre as suas análises e a do especialista. Para isso foram enviados ao jurado os Rótulos de Privacidade, juntamente com suas respectivas políticas textuais. Ambos avaliaram a interpretação do especialista de forma independente um

do outro para assim poderem, depois, confrontar suas classificações. A Tabela 6-6 apresenta a avaliação do jurado.

Tabela 6-6: Avaliação do Júri

	Avaliação	
Rótulo	Júri 1	Júri 2
Política1	Concordo	Concordo
Política2	Concordo	Concordo
Política3	Concordo	Concordo
Política4	Concordo	Concordo
Política5	Concordo	Concordo
Política6	Concordo	Concordo
Política7	Concordo	Concordo
Política8	Concordo	Concordo
Política9	Concordo	Concordo
Política10	Concordo	Concordo

Ficou evidenciado pela Tabela 6-6 que houve 100% de “concordância entre juízes” sobre a interpretação do especialista na geração do Rótulo de Privacidade das políticas utilizadas na feitura da avaliação da proposta deste trabalho, pois ambos estão de acordo.

6.4 Considerações finais

A avaliação conduzida permitiu observar que as políticas de privacidade são compostas por palavras-chaves. Estas palavras são específicas sobre dados dos usuários. De acordo com o levantamento das palavras-chaves associadas aos propósitos de utilização, pôde-se aplicar a técnica desenvolvida para apresentar tais informações aos usuários automaticamente e com certa precisão.

Os resultados evidenciaram que a técnica proposta para casamento de padrões de palavras-chaves de políticas de privacidade foi capaz de apresentar quais dados dos usuários puderam ser coletados e para qual propósito foram

utilizados. Considerando que a técnica fez utilização apenas de palavras-chaves e casamento de padrões, conseguiu-se uma precisão na recuperação das informações de aproximadamente 71%, que, segundo Perez-Castillo et al. (2011), pode ser considerada alta precisão, fornecendo, assim, um nível de confiança considerável.

Também foi avaliada a medida F (*F-Measure*), que indica a exatidão de um classificador ou mecanismo de recuperação de informação. Ela obteve 72%, sendo que, quanto mais próximo de 100%, melhor é o desempenho do mecanismo (GUELPELI, BERNARDINI e GARCIA, 2008; RONCERO, 2010; SEBASTIANI, 2002).

Neste capítulo foram apresentados os passos da metodologia de avaliação do desenvolvimento desta pesquisa, bem como os resultados das métricas para recuperação de informação e avaliação de um júri especializado relativamente à interpretação do autor acerca dos termos das políticas de privacidade analisadas.

Capítulo 7

CONCLUSÕES, LIMITAÇÕES E TRABALHOS FUTUROS

Este capítulo apresenta as conclusões deste trabalho, bem como as contribuições, limitações e trabalhos futuros.

7.1 Conclusões

A maioria dos serviços on-line contém uma política de privacidade. Nestas políticas estão descritos os termos de privacidade com os dados dos usuários, pois às vezes os serviços necessitam coletar suas informações. A forma como os dados são tratados, armazenados e talvez compartilhados com empresas parceiras está descrita nessas políticas.

Porém, segundo alguns autores, por intermédio de estudos com usuários, verificou-se que estes não se preocupam em ler as políticas e, assim, aceitam os termos dos serviços sem mesmo saber quais dados estão sendo coletados e para que serão utilizados. Isso pode ocorrer devido à falta de tempo do usuário, pois o objetivo da utilização do serviço depende da aceitação dos termos. Pode ocorrer até mesmo pelo não entendimento dos termos, que para alguns podem ser complexos e de difícil compreensão (MCDONALD e CRANOR, 2008; MCDONALD et al., 2009; SADEH et al., 2013). Além de as políticas de privacidade conterem termos complexos, também não seguem um padrão de escrita e algumas são extensas, dificultando ainda mais a leitura do usuário.

Pesquisadores elaboraram formas alternativas para a apresentação das políticas de privacidade e uma delas foi a utilização da plataforma P3P, que gera uma tabela contendo as informações sobre coleta e utilização dos dados dos usuários. Os autores fizeram experimentos com usuários e evidenciaram uma

aceitação pelos participantes, podendo, assim, fornecer-lhes informações de forma mais clara.

Como a maioria dos serviços descreve suas políticas em formato textual, pôde-se verificar que, mesmo as políticas não contendo um padrão de escrita, são constituídas de palavras-chaves. Estas palavras se repetem na maioria das políticas, identificando os tipos dos dados dos usuários que podem vir a ser coletados e para quais propósitos são utilizados.

A abordagem proposta permite localizar essas palavras-chaves, criar categorias com agrupamento dessas palavras sobre suas características genéricas, tais como informações pessoais, de contato, localização, entre outras. Com a utilização de casamento de padrões é possível analisar uma particular política de privacidade procurando padrões das categorias. Quando encontrados os padrões de uma determinada categoria, essa informação pode ser apresentada ao usuário de forma mais simples e clara.

Considerando que a maioria dos serviços on-line utiliza linguagem natural em suas políticas de privacidade e que a utilização da apresentação das políticas em formato de tabela pode melhorar o entendimento dos usuários, a abordagem deste trabalho permitiu mapear as categorias elaboradas para uma tabela, denominada de Rótulo de Privacidade. Este Rótulo contém 30 tipos de categorias, nas quais os tipos de informações dos usuários que podem ser coletadas são combinados com os propósitos de serem utilizados.

Para analisar uma particular política de privacidade e mapear as categorias de privacidade para o Rótulo de Privacidade, foi desenvolvido um protótipo de uma aplicação, a PPMark. Com ela é possível analisar uma política e, assim, gerar o Rótulo de forma automática, contendo informações sobre coleta e utilização de dados descritos nos textos. Com a utilização do protótipo para analisar as políticas de privacidade de testes, os resultados obtidos pela abordagem foram consideravelmente precisos nas extrações dos termos de tipos de dados e propósitos de utilização. As contribuições alcançadas neste trabalho são apresentadas a seguir.

7.2 Contribuições e limitações

A principal contribuição deste trabalho foi a elaboração de uma abordagem para encontrar palavras-chaves nas políticas de privacidade. Com estas palavras é possível determinar quais informações dos usuários podem ser coletadas e para que serão utilizadas. Foi feito um agrupamento dessas palavras-chaves criando-se categorias de privacidade que definem tipos de dados e propósitos de utilização. Foram elaboradas regras de casamento de padrões para analisar políticas de privacidade de forma automática e apresentar as informações para os usuários.

Neste trabalho foi adaptada a tabela elaborada por Kelley et al. (2009), na qual a abordagem extrai informações de textos escritos em linguagem natural. Com a utilização das categorias de privacidade e as regras para casamento de padrões, as informações podem ser mapeadas para outros tipos de formato de apresentação.

O protótipo da aplicação PPMark é outra contribuição que permitiu analisar de forma automática uma política de privacidade e apresentar as informações de coleta e utilização de dados no Rótulo de Privacidade. O protótipo com utilização das regras de casamento de padrões mostrou-se precisa na recuperação dessas informações.

As limitações deste trabalho ficam direcionadas para o processamento de linguagem natural, sendo que na abordagem não foram utilizadas técnicas aprofundadas para extrair informações de textos, mas sim um estudo de quais palavras-chaves as políticas contêm, visando a poder definir, por meio da associação entre estas palavras, qual tipo de dado do usuário é coletado e utilizado.

Conforme foi exposto ao longo deste trabalho, a maioria das políticas de privacidade dos serviços on-line é escrita em linguagem natural. A abordagem proposta não pode substituir a leitura e o entendimento dos textos pelo ser humano, porém se mostra eficaz para ser utilizada como suporte para estas ações.

7.3 Trabalhos futuros

Uma vez que os resultados deste estudo mostraram-se consideravelmente precisos, pode ser útil adaptar-se a abordagem para ainda continuar extraindo informações dos textos escritos em linguagem natural, mas direcionando o estudo para pessoas com deficiência visual, com o propósito de possibilitar a eles o acesso a uma aplicação que lhes permitisse ouvir quais termos estão sendo coletados e utilizados. Isso poderia ser de grande ajuda para essas pessoas.

Como a abordagem deste trabalho não foi focada em aprofundar-se na área de Inteligência Artificial, especificamente nas complexas técnicas de processamento de linguagem natural, como trabalho futuro também é interessante desenvolver algum mecanismo que seja capaz de gerar as categorias de privacidade automaticamente. Assim, ao fornecer um corpus de políticas de privacidade, o processo de categorização seria automatizado, se possível sem a intervenção de um especialista.

Além dos trabalhos citados anteriormente, ainda se pode destacar o desenvolvimento de algum mecanismo que seja capaz, por meio das categorias de privacidade, de analisar se realmente o serviço on-line está coletando apenas as informações citadas nas políticas de privacidade, alertando aos usuários que o website não está cumprindo o acordo descrito em sua política.

Capítulo 8

REFERÊNCIAS

ACQUISTI, A.; BRANDIMARTE, L.; LOEWENSTEIN, G. Privacy and human behavior in the age of information. **Science (New York, N.Y.)**, v. 347, n. 6221, p. 509–14, 2015.

ADAMS, J. T. *et al.* **Automated Tracking of Online Service Policies**. Rochester, NY: [s.n.]. Disponível em: <http://papers.ssrn.com/abstract=1989112>. Acesso em: 7 ago. 2015.

ANTÓN, A. I.; EARP, J. B. A requirements taxonomy for reducing Web site privacy vulnerabilities. **Requirements Engineering**, v. 9, n. 3, p. 169–185, 2004.

BLEI, D.; CARIN, L.; DUNSON, D. Probabilistic topic models. **IEEE Signal Processing Magazine**, v. 27, n. 6, p. 55–65, 2010.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. **J. Mach. Learn. Res.**, v. 3, p. 993–1022, 2003.

BOYER, R. S.; MOORE, J. S. A fast string searching algorithm. **Communications of the ACM**, v. 20, n. 10, p. 762–772, 1977.

CANTONE, D.; FARO, S. Fast-search algorithms: New efficient variants of the Boyer-Moore pattern-matching algorithm. **Journal of Automata, Languages and Combinatorics**, v. 10, n. 5/6, p. 589–608, 2005.

CHOUDHARY, A.; ASHAR, P.; KULKARNI, J. **String matching engine,** Google Patents, 2006.

CONGER, S.; PRATT, J. H.; LOCH, K. D. Personal information privacy and emerging technologies. **Information Systems Journal**, v. 23, n. 5, p. 401–417, 2013.

COSTANTE, E. *et al.* **A machine learning solution to assess privacy policy completeness.** In: PROCEEDINGS OF THE 2012 ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY - WPES '12. **Anais**...: WPES '12.New York, NY, USA: ACM, 2012Disponível em: <http://dl.acm.org/citation.cfm?id=2381966.2381979>

FEDERAL TRADE COMMISSION. Privacy online : fair informaion practices in the electronic marketplace. **Statement of the Federal Trade Commission before the Committee on Commerce, Science and Transportation, United States Senate, Washington, DC**, p. 208, 2000.

FERRAIOLO, D. F. *et al.* Proposed NIST standard for role-based access control. **ACM Transactions on Information and System Security**, v. 4, n. 3, p. 224–274, 2001.

FONSECA, R.; SILVA, P.; SILVA, R. Acordo inter-juízes: O caso do coeficiente kappa. **Laboratório de Psicologia**, v. 5, n. 1, p. 81–90, 2007. Disponível em: <http://publicacoes.ispa.pt/index.php/lp/article/view/759/702>

GHANI, N. A.; SIDEK, Z. M. **Controlling your personal information disclosure.** In: PROCEEDINGS OF THE 7TH WSEAS INTERNATIONAL CONFERENCE ON INFORMATION SECURITY AND PRIVACY. **Anais**...2008

GUELPELI, M. V. C.; BERNARDINI, F. C.; GARCIA, A. C. B. Todas as palavras da sentença como métrica para um sumarizador automático. In: COMPANION PROCEEDINGS OF THE XIV BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB - WEBMEDIA '08, p. 287, 2008.

JENSEN, C.; POTTS, C. **Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices.** In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS. **Anais**...2004

KALSI, P.; PELTOLA, H.; TARHIO, J. **Comparison of Exact String Matching**. [s.l.] Springer, 2008.

KARP, R. M.; RABIN, M. O. Efficient randomized pattern-matching algorithms. **IBM Journal of Research and Development**, v. 31, n. 2, p. 249–260, 1987.

KELLEY, P. G. *et al.* **A "nutrition label" for privacy.** In: PROCEEDINGS OF THE 5TH SYMPOSIUM ON USABLE PRIVACY AND SECURITY SOUPS 09. **Anais**...2009. Disponível em: <http://portal.acm.org/citation.cfm?doid=1572532.1572538>

___. **Standardizing Privacy Notices: An Outline Study of the Nutritional Label Approach.** In: PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS. **Anais**...2010

LANDGREBE, T. C. W. *et al.* Precision-Recall Operating Characteristic (P-ROC) curves in imprecise environments. In: PROCEEDINGS - INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, v. 4, p. 123–127, 2006.

LIU, L.; ÖZSU, M. T. **Encyclopedia of Database Systems**. [s.l.] Springer Publishing Company, Incorporated, 2009.

LOBATO, L. L.; ZORZO, S. D. Padrões para apoio ao desenvolvimento de Políticas de Privacidade. **Supporting Organizations**, p. 3, 2007a.

___. Avaliação dos Mecanismos de Privacidade e Personalização na Web Universidade Federal de São Carlos Avaliação dos Mecanismos de Privacidade e Personalização na Web. **Universidade Federal de São Carlos, São Paulo**, 2007b.

MASSEY, A. K. *et al.* Automated text mining for requirements analysis of policy documents. In: PROCEEDINGS - 21ST IEEE INTERNATIONAL REQUIREMENTS

ENGINEERING CONFERENCE, RE 2013. **Anais**...2013

MATOS, D. A. S. Confiabilidade e concordância entre juízes : aplicações na área educacional. **Estudos em Avaliação Educacional**, v. 25, n. 59, p. 298–324, 2014.

MCDONALD, A.; CRANOR, L. F. The Cost of Reading Privacy Policies. **I/S - A Journal of Law and Policy for the Information Society**, v. 4, n. 3, p. 1–22, 2008.

MCDONALD, A. M. *et al.* **A Comparative Study of Online Privacy Policies and Formats.** In: 9TH INTERNATIONAL SYMPOSIUM ON PRIVACY ENHANCING TECHNOLOGIES. **Anais**...2009. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03168-7_3\nhttp://link.springer.com/10.1007/978-3-642-03168-7_3>

MEDEIROS VANDERLEI, I. Casamento de Padrão em Strings Privados , com Aplicação em Consultas Seguras a Banco de Dados. **Universidade Federal de Pernambuco**, 2006.

MIKOLOV, T. *et al.* Distributed Representations of Words and Phrases and their Compositionality. **Nips**, p. 1–9, 2013.

MOREIRA, G. DE A. Algoritmos para Busca de Padrões: Uma Análise Comparativa Empírica. **Journal of Chemical Information and Modeling**, p. 1689–1699, 2012.

NEBEL, M. E. Fast string matching by using probabilities: On an optimal mismatch variant of Horspool's algorithm. **Theoretical Computer Science**, v. 359, n. 1-3, p. 329–343, 2006.

NEFF, M. S.; BYRD, R. J.; BOGURAEV, B. K. The Talent System: TEXTRACT Architexture and Data Model. In: PROCEEDINGS OF THE HLT-NAACL 2003 WORKSHOP ON SOFTWARE ENGINEERING AND ARCHITECTURE OF LANGUAGE TECHNOLOGY SYSTEMS - SEALTS '03, v. 10, n. 3-4, p. 1–8, 2003.

PANDITA, R. Inferring Semantic Information from Natural-Language Software Artifacts. **North Carolina State University**, p. 1–31, 2013. Disponível em: < http://rahulpandita.me/files/Prelim_Report.pdf>

PEARSON, S. Taking Account of Privacy when Designing Cloud Computing Services 2. Why is it important to take privacy into…2009. In: CHALLENGES OF CLOUD COMPUTING, 2009. CLOUD' …. **Anais**...2009. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5071532>

PEREZ-CASTILLO, R. *et al.* Obtaining Thresholds for the Effectiveness of Business Process Mining. In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT (ESEM), 2011. **Anais**...2011. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6092604&matchBoolean=true&pageNumber=2&searchField=Search_All&queryText=(((p_Publication_Title:"Process+mining")+OR+p_Abstract:"Process+mining")+)\nhttp://dx.doi.org/10.1109/ESEM.2011.64>

RAMOS, J.; EDEN, J.; EDU, R. Using TF-IDF to Determine Word Relevance in Document Queries Processing. In: PROCEEDINGS OF THE FIRST INSTRUCTIONAL CONFERENCE ON MACHINE LEARNING. **Anais**...2003

Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf>

REIDENBERG, J. R. *et al.* Disagreeable privacy policies: Mismatches between meaning and users' understanding. **Berkeley Technology Law Journal**, v. 30, n. 1, p. 39–68, 2015.

RONCERO, V. G. **Classificação semi-supervisionada de textos em ambientes distribuídos**. [s.l.] Universidade Federal do Rio de Janeiro, 2010.

SADEH, N. *et al.* **The Usable Privacy Policy Project**. Carnegie Mellon University, 2013. (Technical Report, CMU-ISR-13-119).

SARDINHA, T. Corpus Linguistics: history and problematization. **DELTA: Documentação de Estudos em Lingüística …**, v. 16, n. n.2, p. 323–367, 2000.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys**, v. 34, n. 1, p. 1–47, 2002.

SINGLA, N.; GARG, D. String Matching Algorithms and their Applicability in various Applications. **International Journal of Soft Computing and Engineering**, v. 1, n. 6, p. 218–222, 2012.

SMIT, E. G.; NOORT, G. VAN; VOORVELD, H. A. M. Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe. **Computers in Human Behavior**, v. 32, n. 0, p. 15–22, 2014.

SMITH, R.; XU, J. A survey of personal privacy protection in public service mashups. In: PROCEEDINGS - 6TH IEEE INTERNATIONAL SYMPOSIUM ON SERVICE-ORIENTED SYSTEM ENGINEERING, SOSE 2011. **Anais**...2011

SOLOVE, D. J. A Taxonomy of Privacy. **University of Pennsylvania Law Review**, v. 154, n. 3, p. 477–560, 2006.

SPÄRCK JONES, K. A Statistical Interpretation of Term Specificity and its Retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972.

SUSIK, R.; GRABOWSKI, S.; FREDRIKSSON, K. Multiple pattern matching revisited. **Cornell University Library**., v. abs/1405.5, 2014. Disponível em: < http://arxiv.org/pdf/1405.5483.pdf>

SUSTIK, M.; MOORE, J. **String searching over small alphabets**. [s.l.] Computer Science Department, University of Texas at Austin, 2007.

WAIN, K. *et al.* PScout : Analyzing the Android Permission Specification. In: CCS '12 PROCEEDINGS OF THE 2012 ACM CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY. **Anais**...2012Disponível em: <http://www.eecg.toronto.edu/~lie/papers/PScout-CCS2012-web.pdf\nhttp://dl.acm.org/citation.cfm?id=2382222>

WATSON, B. W. A new regular grammar pattern matching algorithm. Lecture Notes

in: Computer Science, Springer, v. 1136, n. 1, p. 364–377, 1996.

WESTIN, A. F. Privacy and Freedom. **American Sociological Review**, v. 33, n. 1, p. 173, 1968.

XIAO, X.; PARADKAR, A.; XIE, T. Automated extraction and validation of security policies from natural-language documents**.** Perspective. In: FSE'12, 2011, New York. **Anais**... NY: ACM, 2011. Disponível em: <http://domino.research.ibm.com/library/cyberdig.nsf/papers/25F751FE19E2E98F85257871004CA1FD/$File/rc25128.pdf>

ZORZO, S. D.; LOBATO, L. L. Avaliação por Inspeção em Sites Brasileiros de ecommerce: um Estudo de Caso. Relatório Técnico. São Carlos: Departamento de Computação da Universidade Federal de São Carlos, 2007.

... uma nova temporada se inicia aqui.

Diego Roberto Gonçalves de Pontes