UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CÁSSIO ALAN GARCIA

Extração de Informações de Conferências em Páginas Web

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Viviane P. Moreira

Porto Alegre
2017

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Garcia, Cássio Alan

Extração de Informações de Conferências em Páginas Web / Cássio Alan Garcia. – Porto Alegre: PPGC da UFRGS, 2017.

50 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Orientador: Viviane P. Moreira.

1. Extração de informação. 2. Conferências. 3. Deadlines. 4. Conditional Random Fields. I. Moreira, Viviane P.. II. Título.

“If I have seen farther than others,
it is because I stood on the shoulders of giants.”
— Sir Isaac Newton

AGRADECIMENTOS

Agradeço inicialmente por ter a oportunidade de cursar o Mestrado, tem sido um grande aprendizado, muito além do âmbito acadêmico, mas também como experiência de vida. Ainda mais que a graduação e o trabalho, o Mestrado deixou bastante claro o significado de disciplina, foco e também a necessidade de abrir mão de muitas atividades ao lutar por um objetivo. Neste sentido, agradeço a compreensão dos amigos próximos e da família pelos numerosos momentos de ausência, mesmo assim obtendo constante apoio e torcida para o sucesso. Agradeço em especial a meus pais Francisco e Valmira, e irmãos por todo o incentivo durante essa jornada. Foram sempre exemplo de que é possível crescer, almejar objetivos cada vez maiores e, com esforço e dedicação, alcançá-los. Agradecimento aos amigos colegas que contribuíram com questionamentos e opiniões. Agradeço aos professores envolvidos na minha formação, principalmente a professora Dra. Viviane Pereira Moreira, por ter me dado a oportunidade de ser seu orientando, indicando o caminho a tomar, com sugestões valiosas a todo momento, e estando constantemente disponível, demonstrando toda a atenção.

RESUMO

A escolha da conferência adequada para o envio de um artigo é uma tarefa que depende de diversos fatores: (i) o tema do trabalho deve estar entre os temas de interesse do evento; (ii) o prazo de submissão do evento deve ser compatível com tempo necessário para a escrita do artigo; (iii) localização da conferência e valores de inscrição são levados em consideração; e (iv) a qualidade da conferência (Qualis) avaliada pela CAPES. Esses fatores aliados à existência de milhares de conferências tornam a busca pelo evento adequado bastante demorada, em especial quando se está pesquisando em uma área nova. A fim de auxiliar os pesquisadores na busca de conferências, o trabalho aqui desenvolvido apresenta um método para a coleta e extração de dados de sites de conferências. Essa é uma tarefa desafiadora, principalmente porque cada conferência possui seu próprio site, com diferentes layouts. O presente trabalho apresenta um método chamado CONFTRACKER que combina a identificação de URLs de conferências da Tabela Qualis à identificação de deadlines a partir de seus sites. A extração das informações é realizada independente da conferência, do layout do site e da forma como são apresentadas as datas (formatação e rótulos). Para avaliar o método proposto, foram realizados experimentos com dados reais de conferências da Ciência da Computação. Os resultados mostraram que CONFTRACKER obteve resultados significativamente melhores em relação a um *baseline* baseado na posição entre rótulos e datas. Por fim, o processo de extração é executado para todas as conferências da Tabela Qualis e os dados coletados populam uma base de dados que pode ser consultada através de uma interface online.

Palavras-chave: Extração de informação. Conferências. Deadlines. Conditional Random Fields.

Information Extraction from Conference Web Pages

ABSTRACT

Choosing the most suitable conference to submit a paper is a task that depends on various factors: (i) the topic of the paper needs to be among the topics of interest of the conference; (ii) submission deadlines need to be compatible with the necessary time for paper writing; (iii) conference location and registration costs; and (iv) the quality or impact of the conference. These factors allied to the existence of thousands of conferences, make the search of the right event very time consuming, especially when researching in a new area. Intending to help researchers finding conferences, this work presents a method developed to retrieve and extract data from conference web sites. Our method combines the identification of conference URL and deadline extraction. This is a challenging task as each web site has its own layout. Here, we propose CONFTRACKER, which combines the identification of the URLs of conferences listed in the Qualis Table and the extraction of their deadlines. Information extraction is carried out independent from the page's layout and how the dates are presented. To evaluate our proposed method, we carried out experiments with real web data from Computer Science conferences. The results show that CONFTRACKER outperformed a baseline method based on the position of labels and dates. Finaly, the extracted data is stored in a database to be searched with an online tool.

Keywords: Information Extraction, Conditional Random Fields.

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CSS	*Cascading Style Sheets*
CFP	*Call for Papers*
CRF	*Conditional Random Fields*
DBLP	*Digital Bibliography and Library Project*
DOM	*Document Object Model*
HMM	*Hidden Markov Model*
HTML	*HyperText Markup Language*
NER	*Named Entity Recognition*
OCR	*Optical Character Recognition*
POS	*Part of Speech*
SVM	*Suport Vector Machine*
URL	*Uniform Resource Locator*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O processo de escrita e submissão de artigos científicos é crucial na vida dos pesquisadores. A escolha do periódico ou conferência mais adequados para a divulgação da pesquisa realizada é uma tarefa bastante importante e que por vezes demanda bastante tempo dos pesquisadores, sendo necessário acessar o site de cada uma das conferências de interesse, para então localizar e levantar as informações relevantes para a escolha da mais adequada.

Existem milhares de conferências científicas que ocorrem anualmente. Quando se deseja submeter um artigo para uma conferência, vários aspectos precisam ser levados em consideração: (i) tema do trabalho deve estar entre os temas de interesse do evento para que ele possa ser considerado; (ii) necessidade em saber se os prazos (*deadlines*) do evento são compatíveis com os do término da escrita do artigo (ou algum outro critério temporal como o prazo para a conclusão do curso, por exemplo); (iii) questões de valores financeiros como o local de realização da conferência e o valor da taxa de inscrição que podem inviabilizar a participação dos autores; e (iv) a qualidade do evento também é importante para essa escolha – um trabalho em nível inicial pode ser enviado para um evento de menor impacto, enquanto que um trabalho de final de doutorado, por exemplo, pode ser enviado para um evento de maior qualificação.

No Brasil, a avaliação da produção intelectual de programas de pós-graduação é feita pela CAPES e baseia-se no sistema Qualis. O Qualis (SOUZA; PAULA, 2002) é um instrumento avaliativo cujo resultado é uma tabela que atribui um grau a conferências e periódicos que visa refletir sua qualidade. Os graus são chamados *estratos* e podem assumir os valores A1, A2, B1, B2, B3, B4, B5 e C, sendo A1 o mais elevado.

A tarefa de encontrar uma conferência que seja adequada ao tema, ao prazo e ao Qualis costuma consumir um tempo considerável dos pesquisadores, em especial quando se está começando em uma área de pesquisa nova da qual não se conhece os eventos. Existem alguns web sites que reúnem chamadas de submissão de artigos (*call for papers - CFPs*), tais como ConfSearch[1] e WikiCFP[2], contudo a maioria se baseia na inclusão manual de informações. O objetivo desse trabalho é preencher essas lacunas ao buscar de forma automatizada essas informações, permitindo que a comunidade possa consultar dados atualizados sobre CFPs.

[1] <http://www.confsearch.org/confsearch/>
[2] <http://wikicfp.com/cfp/>

1.1 Objetivo

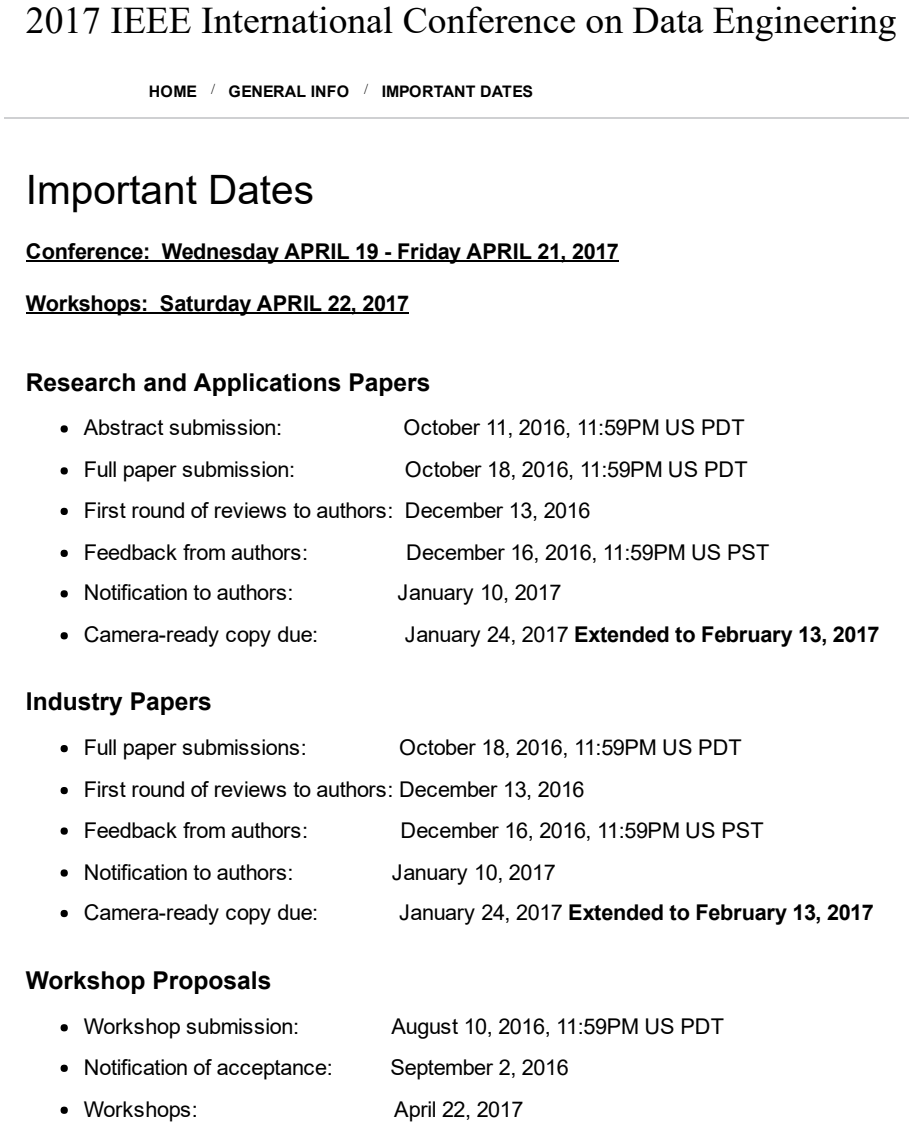
Nesse trabalho é proposto o CONFTRACKER, o qual apresenta um método automático de coleta e extração de dados das conferências listadas na Tabela Qualis. Os dados extraídos são armazenados em um banco de dados de eventos, o qual é disponibilizado por meio de um ambiente próprio de pesquisa, que permite a busca por palavras-chaves, prazos de submissão e Qualis das conferências.

Então, a partir da Tabela Qualis, é localizado o site de cada uma das conferências, feito o download do conteúdo Web, para, aplicando técnicas de Aprendizado de Máquina, mais especificamente a de *Conditional Random Fields - CRF*, extrair as datas dos principais *deadlines*, armazenando e disponibilizando-os por meio da ferramenta de pesquisa própria.

A tarefa de extração de dados de conferências é desafiadora por uma série de motivos:

- Layout das páginas de conferência: cada evento possui seu próprio site, dispondo das informações de maneiras diferentes. Sendo assim, não há como desenvolver um *template* de extração de forma que se encontre as datas de interesse em sua exata disposição na página. Devido ao grande número de conferências, também é impraticável desenvolver o template para cada evento especificamente. Ainda, mesmo que o número fosse menor, com o passar do tempo o layout de uma página vai sendo atualizado e alterado, fazendo com que uma abordagem inflexível seja nada vantajosa.
- Uso de linguagem natural: para cada data presenta nas páginas de *deadlines*, existem diversas formas de serem rotuladas. A exemplo, a data de submissão de resumo pode estar descrita como "*abstract submission*", ou "*paper abstract due*", entre outros.
- Correta associação entre data e rótulo: é necessário vincular corretamente as datas encontradas a seus rótulos, de forma que a data do *deadline* de submissão de artigo não seja atribuída ao *deadline* de submissão de resumo, por exemplo.
- Atualização das datas: existem casos em que os *deadlines* de submissão de artigos são postergados próximo a data divulgada inicialmente (Figura 1.1). Sendo assim, é necessário ser refeita a extração das datas da conferência em questão a fim de verificar se houve alteração e atualizar na base de dados.

Figura 1.1: Várias *tracks* em uma mesma conferência. Extraída da página do ICDE 2017 http://icde2017.sdsc.edu/dates



2017 IEEE International Conference on Data Engineering

HOME / GENERAL INFO / IMPORTANT DATES

Important Dates

Conference: Wednesday APRIL 19 - Friday APRIL 21, 2017

Workshops: Saturday APRIL 22, 2017

Research and Applications Papers

- Abstract submission: October 11, 2016, 11:59PM US PDT
- Full paper submission: October 18, 2016, 11:59PM US PDT
- First round of reviews to authors: December 13, 2016
- Feedback from authors: December 16, 2016, 11:59PM US PST
- Notification to authors: January 10, 2017
- Camera-ready copy due: January 24, 2017 **Extended to February 13, 2017**

Industry Papers

- Full paper submissions: October 18, 2016, 11:59PM US PDT
- First round of reviews to authors: December 13, 2016
- Feedback from authors: December 16, 2016, 11:59PM US PST
- Notification to authors: January 10, 2017
- Camera-ready copy due: January 24, 2017 **Extended to February 13, 2017**

Workshop Proposals

- Workshop submission: August 10, 2016, 11:59PM US PDT
- Notification of acceptance: September 2, 2016
- Workshops: April 22, 2017

Fonte: Adaptado de http://icde2017.sdsc.edu/dates

1.2 Contribuições

As principais contribuições desde trabalho podem ser sumarizadas como:

- a proposta de um método para a extração de datas de interesse em sites de conferências e;
- a disponibilização de um site <http://inf.ufrgs.br/conftracker/> que reúne as datas das conferências listadas na Tabela Qualis da Computação.

1.3 Organização

O restante desse trabalho está organizado da seguinte forma: o Capítulo 2 introduz conceitos básicos e o Capítulo 3 trabalhos relacionados em se tratando de Extração de Informações na Web. O Capítulo 4 mostra uma visão geral sobre a solução proposta, enquanto o Capítulo 5 explica detalhadamente a etapa de extração de datas das conferências. No Capítulo 6 são apresentados experimentos e avaliações realizadas e por fim o Capítulo 7 apresenta as conclusões e trabalhos futuros.

2 REFERENCIAL TEÓRICO

A transformação de dados não estruturados disponíveis em páginas Web em dados estruturados facilita a localização da informação. Enquanto em páginas web a busca se faz por *keywords* ou pesquisas seguindo *links*, em uma base de dados estruturada é possível localizar diretamente, por meio de uma consulta SQL, por exemplo, a informação sendo buscada. Sendo uma tarefa vastamente pesquisada inicialmente na área de processamento de linguagem natural, com o passar do tempo, técnicas de diversas áreas são propostas na literatura endereçando esse problema de transformação de dados não estruturados em estruturados, tais como das áreas de aprendizado de máquina, base de dados e ontologias (LAENDER et al., 2002; SARAWAGI et al., 2008).

Nesse capítulo são abordadas duas técnicas de Aprendizado de Máquina. Na Seção 2.1, é dada uma introdução às Cadeias de Markov, necessária para o entendimento do que segue na Seção 2.2, sobre *Conditional Random Fields*, técnica utilizada para a solução proposta nesse trabalho.

2.1 Modelos Ocultos de Markov

O Modelo Oculto de Markov (ou *Hidden Markov Model - HMM*) é uma técnica utilizada para rotular (explicar e caracterizar) uma sequência de símbolos, podendo esses serem discretos, contendo um número finito de estados, ou contínuos. Dado que observações geradas por problemas do mundo real podem facilmente ser representadas como uma sequência de símbolos, esse modelo pode ser aplicado a tarefas relacionadas a rotulação de dados (RABINER; JUANG, 1986; RABINER, 1989).

Um exemplo simples colocado por Rabiner e Juang (1986) é o do jogo da moeda, cara ou coroa. Nesse exemplo, é realizado um experimento jogando várias vezes a moeda, chegando à seguinte observação, sendo que H representa cara, e T coroa:

$$O = H\ H\ T\ H\ T\ T\ H\ H\ \ldots\ T$$

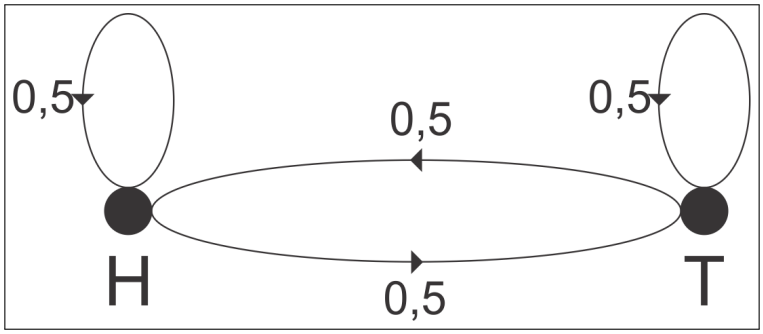
Dado esse experimento, pode-se construir uma cadeia de Markov que represente a sequência observada. Um modelo possível está representado na Figura 2.1, onde há dois estados (nodos), representando os dois lados da moeda, e as arestas colocando a relação entre os estados, mostrando a probabilidade da troca de um estado para outro, ou

de manter-se.

Cadeias de Markov modelam as transições de estado de processos em que o próximo estado dependa exclusivamente do estado corrente, ou seja, dado o presente, o futuro independe do passado. Essa característica de "não-memória" é conhecida como a *propriedade de Markov*. Um processo randômico define um sistema em que em um dado tempo t está em um determinado estado e a transição de estado ocorre randomicamente.

A Figura 2.1 é um exemplo de representação de uma cadeia de Markov, sendo um grafo direcionado em que as arestas são ponderadas em função da probabilidade de um estado mudar para outro, esses sendo representados pelos nodos.

Figura 2.1: Cadeia de Markov representando estados de uma moeda



Fonte: Adaptado de Rabiner e Juang (1986)

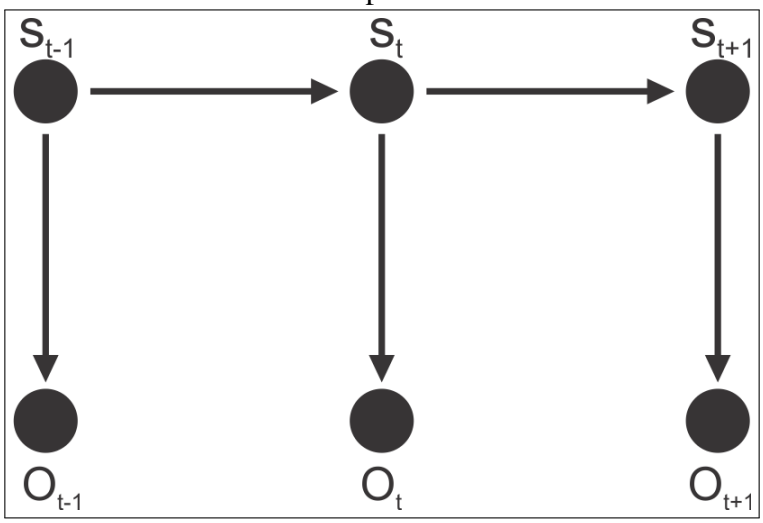
As limitações desse modelo vêm do fato de que problemas do mundo real dificilmente podem ter todos seus estados observados, como no caso do exemplo do cara-coroa. Existem informações que podem não estar sendo observadas. Partindo deste princípio, passa-se a dividir os estados entre observáveis e ocultos.

Os Modelos Ocultos de Markov (HMM) são Modelos de Markov em dois estágios (Figura 2.2), sendo o primeiro deles um processo randômico discreto com estados finitos. No segundo estágio, para cada instante de tempo t uma emissão O_t é gerada (FINK, 2014; RABINER; JUANG, 1986; RABINER, 1989).

A distribuição de probabilidade para as emissões são dependentes apenas do estado atual S_t, não dependendo de estados ou emissões anteriores. Na Figura 2.2, a aresta ligando um estado S a outro é a mesma das cadeias de Markov. As arestas ligando um estado S a O, representam a probabilidade da emissão.

A principal limitação dos Modelos Ocultos de Markov é assumir-se que um elemento da observação depende apenas do estado naquele momento. Ao rotular dados sequenciais, a observação anterior possui forte influência sobre o rótulo da próxima observação. A exemplo, em tarefas de rotulação morfossintática, ou reconhecimento de entidades nomeadas, a palavra anterior pode influenciar fortemente na rotulação da pró-

Figura 2.2: Cadeia de Markov representando estados de uma moeda



Fonte: Adaptado de Fink (2014)

xima palavra (LAFFERTY et al., 2001). Para sanar esse problema, Lafferty et al. (2001) introduzem a técnica de *Conditional Random Fields*, detalhado na seção 2.2.

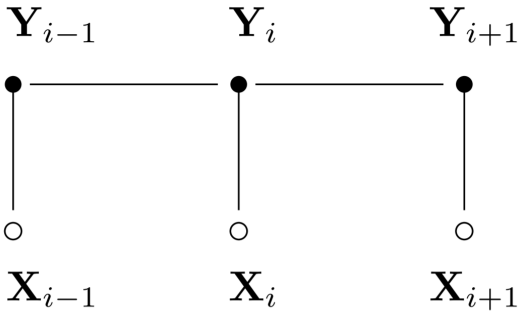
2.2 Conditional Random Fields - CRF

Conditional Random Fields é um *framework* para construção de modelos probabilísticos para segmentação e rotulação de sequência de dados (LAFFERTY et al., 2001). É uma técnica amplamente utilizada na área de extração de informação e mostrando ótimos resultados, conforme exposto no Capítulo 3. Nessa seção é feita uma introdução a esse *framework* utilizado na etapa de extração de dados.

Na Figura 2.3, X é uma variável randômica dentre a sequência de dados a serem rotulados, e Y uma variável randômica sobre a sequência de *rótulos*. Todo $Y_i \in Y$ pode assumir valores de um alfabeto finito A. Por exemplo, X pode assumir valores de uma sentença em linguagem natural, enquanto Y assume valores de *rótulos* de *Part-of-Speech - POS* (Etiquetador Morfo Sintático) daquelas sentenças (LAFFERTY et al., 2001).

Segundo Wallach (2004), um CRF pode ser visto como um modelo em forma de grafo não direcionado condicionado em X, sendo X a variável randômica representando a sequência sendo rotulada. Definindo formalmente, dado o grafo $G = (V, E)$, em que $Y = (Y_v)_{v \in V}$, sendo Y os vértices de G. Então (X, Y) é um CRF quando, condicionada a X, a varável randômica Y_v obedece a propriedade de Markov respeitando o grafo $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, onde $w \sim v$ significa que w e v são vizinhos

Figura 2.3: Representação de um modelo CRF. Variáveis randômicas X não são geradas pelo modelo



Fonte: Adaptado de Lafferty et al. (2001)

em G (LAFFERTY et al., 2001; WALLACH, 2004).

A fórmula a seguir define a probabilidade de uma sequência de *rótulos* y dada uma sequência de observação x para ser um produto normalizado de funções potenciais:

$$p(Y|X) = exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)) \tag{2.1}$$

onde $\lambda_j t_j(y_{i-1}, y_i, x, i)$ representa a *feature* de transição da sequência de observação todos os *rótulos* em posições i e $i-1$ na sequência de *rótulos*; $s_k(y_i, x, i)$ é a *feature* de estado do rótulo na posição i e a sequência de observação; e λ_j e μ_k são parâmetros a serem estimados com dados de treinamento.

Formalmente definindo os parâmetros λ_j e μ_k:

$$\theta = (\lambda_1, \lambda_2, ...; \mu 1, \mu 2, ...)$$

Assumindo os dados de treinamento com $(x^{(k)}, y^{(k)})$, o produto da Equação 2.1 sobre toda a sequência de treinamento como a função de θ é conhecida como verossimilhança (*likelihood*), definido por $p(y^{(k)}, x^{(k)}, \theta)$. O treinamento para a máxima verossimilhança (*maximum likelihood*) busca valores de parâmetros que maximizam o logaritmo da verossimilhança (LAFFERTY et al., 2001; WALLACH, 2004).

Ao definir *features* se constrói um conjunto de valores $b(x, i)$ para expressar alguma característica da distribuição empírica dos dados de treinamento que também vem a ser a distribuição do modelo (WALLACH, 2004):

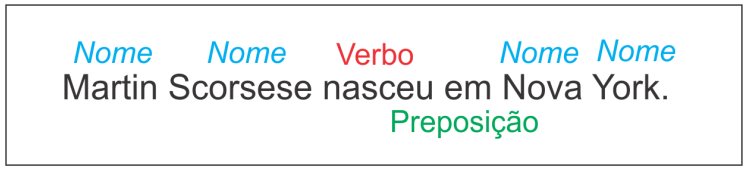
$$b(x,i) = \begin{cases} 1 & \text{se a observação na posição } i \text{ é uma palavra com inicial maiúscula} \\ 0 & \text{caso contrário} \end{cases}$$

Da mesma forma, pode-se definir *feature* de transição:

$$t_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x,i) & \text{se } y_{i-1} = \text{inicia maiúscula e } y_i = \text{numero de letras } > n \\ 0 & \text{caso contrário} \end{cases}$$

Na Figura 2.4 é apresentado um exemplo de tarefa de rotulação morfossintática (*Part of Speech - POS*), em que na frase "Martin Scorsese nasceu em Nova York", a tarefa é identificar as diferentes funções sintáticas das palavras na frase. Os rótulos estão destacados na Figura.

Figura 2.4: Exemplo de rotulação de POS



Fonte: O Autor

Na Tabela 2.1, a primeira coluna refere-se aos *tokens* deste exemplo, a segunda representa um exemplo de *feature* criada (nesse caso, indicando se a palavra começa com letra maiúscula) e a última coluna sendo o rótulo a ser ensinado ao CRF no momento do treinamento e o qual deve ser predito pelo CRF no momento da aplicação do modelo gerado. Para prever qual o rótulo atribuir para cada *token*, o CRF leva em consideração as *features* definidas.

Tabela 2.1: Dados para treinamento do CRF para realização de tarefa de POS

Token	Primeira Maiúscula	Rótulo
Martin	1	Nome
Scorsese	1	Nome
nasceu	0	Verbo
em	0	Preposição
Nova	1	Nome
York	1	Nome

Fonte: O autor

2.3 Sumário

Nesse capítulo abordou-se técnicas pesquisadas para a aplicação nesse trabalho. O Modelo Oculto de Markov (HMM) possui suas limitações principalmente por desconsiderar observações próximas ao dado sendo rotulado. Entendendo-se o funcionamento do HMM, pode-se apresentar o *framework Conditional Random Fields*, proposto por Lafferty et al. (2001), o qual trabalha também com observações próximas ao dado sendo rotulado.

O CRF vem sendo amplamente utilizado em trabalhos na área de extração de informação, obtendo bons resultados. Essa técnica foi a escolhida para a tarefa de extração das datas das conferências, detalhada no Capítulo 5, principalmente devido a sua característica de trabalhar com informações próximas ao dado sendo rotulado: para extrair corretamente a data associada a cada um dos *deadlines*, se faz necessário analisar os rótulos que se encontram em suas proximidades.

3 TRABALHOS RELACIONADOS

Nesse capítulo, são apresentados os principais trabalhos relacionados ao tema dessa dissertação. Assim, as próximas seções desse Capítulo estão organizadas da seguinte forma: na Seção 3.1 serão abordados os trabalhos que visam extrair dados de conferências especificamente; na Seção 3.2 serão abordados os trabalhos que tratam da extração de informações de páginas Web com o uso específico de CRF; e na Seção 3.3, trabalhos que visam a extração de dados não-estruturados em páginas Web com a utilização de outras técnicas.

3.1 Repositórios e extratores de datas de conferências

Durante pesquisas, verificou-se a existência de trabalhos que centralizam informações referentes a conferências. Por exemplo, o *WikiCFP*[1] é um repositório que reúne milhares de CFPs para eventos nas áreas de ciência e tecnologia. O site relata receber cerca de 100 mil visitas mensais. Contudo, o site depende que os dados das CFPs sejam inseridos manualmente por um usuário (por exemplo, um dos organizadores do evento).

Trabalhando da mesma forma, o *Conference Alarm*[2] aparece como alternativa, permitindo a pesquisa por conferências previamente cadastradas de forma manual por um usuário, porém exibe apenas as datas do evento, sem os *deadlines* de submissão.

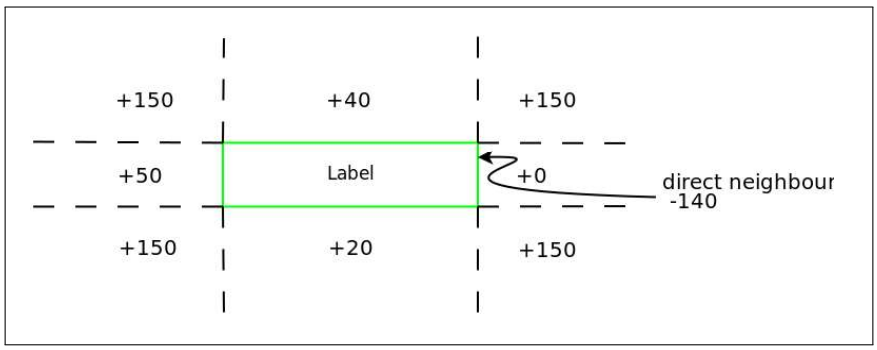
Por sua vez, o *AllCall* (CORREIA et al., 2010), mais automatizado, é um sistema que processa CFPs contidas em e-mails a fim de extrair tópicos e datas importantes. Para que o sistema funcione, o usuário precisa estar inscrito em listas de e-mails que recebam mensagens com CFPs. Dividido em três módulos, (i) o sistema inicialmente conecta com a caixa de entrada do *e-mail* e faz a extração do corpo dos *e-mails*. Em seguida (ii) os *e-mails* são filtrados mantendo-se apenas os que tenham as seguintes palavras de interesse: "*Conference*", "*Symposium*", "*Colloquium*", "*Meeting*" ou "*Workshop*", os quais passam pela extração das datas. A extração no *AllCall* acontece analisando-se o texto, em que ao se encontrar um termo de interesse, do tipo "*Submission Deadline*", verifica-se à direita e esquerda se existe algum padrão de datas, para que esse seja associado ao *deadline* de Submissão do Artigo (nesse exemplo) e persistido em uma base de dados. Extrai-se também a localização do evento, verificando se país e cidade encontram-se dentre os pre-

[1] <http://wikicfp.com/cfp/>
[2] <http://conferencealarm.com/>

viamente cadastrados no *AllCall*. Por fim, utiliza-se padrões para a extração dos tópicos da conferência, iniciando a busca onde for encontrado o termo “Topics"e parando quando encontrar algumas palavras reservadas observadas na maioria dos CFPs analizados. O terceiro módulo (iii) trata-se de um ambiente web para o usuário realizar consultas às informações extraídas e salvas na base de dados.

Já o *ConfSearch*[3] baseia-se em dados do DBLP, um importante repositório de dados bibliográficos da Ciência da Computação. Dados de conferências que não estão disponíveis no DBLP[4], como os prazos de submissão, precisam ser cadastrados pelos usuários do site. Como uma forma de automatizar esse processo manual, Mattes (2011) propõe uma forma de localizar as datas de interesse avaliando a disposição visual das informações ao longo de páginas Web. Quando um usuário deseja cadastrar uma nova conferência, ou complementar alguma importada do DBLP com informações não fornecidas por esse, necessita preencher ao menos treze campos, com datas em formatos específicos. No que o trabalho propõe, o usuário fornece apenas a URL da conferência, para que as datas sejam extraídas automaticamente. A solução proposta para a realização da extração das datas é a análise posicional dos elementos na página renderizada. É feita a relação entre datas encontradas na página e rótulos que estejam relacionados às datas sendo extraídas: *deadline* de submissão de resumo, artigo, data da notificação de aceitação, *deadline* do *camera-ready* e datas da conferência. A relação entre rótulo e data é feita medindo a distância entre as bordas mais próximas entre eles, avaliando também a posição em que se encontram (lado a lado, abaixo, na diagonal, etc.). A Figura 3.1 mostra os valores que o autor definiu para esse acréscimo de valor da distância em função do posicionamento entre o rótulo e data.

Figura 3.1: ConfSearch - acréscimo de valor de distância em função do posicionamento entre nodos



Fonte: Mattes (2011)

[3]<http://www.confsearch.org/confsearch/>
[4]<http://dblp.uni-trier.de/>

Então, para cada *deadline* a ser extraído, escolhe-se a data com a menor distância total, ignorando caso esse valor for maior que um dado limite, sendo este o caso de não ter sido encontrada data para aquele *deadline*. Verifica-se também a consistência entre as datas extraídas, ou seja, se estão em ordem e não há grandes intervalos entre elas.

O CONFTRACKER, proposto nesse trabalho, difere do *WikiCFP*, do *Conference Alarm* e do *ConfSearch* por fazer a coleta e a extração dos dados das conferências de forma automática. Em relação ao *AllCall*, nosso diferencial está em não depender do recebimento de emails com CFPs e liberação do usuário para acesso aos seus emails. Além disso, nenhum dos sistemas pesquisados leva em conta a classificação Qualis dos eventos – que, no Brasil, tem importância.

3.2 Extração de Informações com *Conditional Random Fields*

Conditional Random Fields (CRF), conforme descrito na Seção 2.2, é um modelo estocástico utilizado habitualmente para etiquetar e segmentar sequências de dados ou extrair informações de documentos textuais (LAFFERTY et al., 2001).

Uma das aplicações do CRF concentra-se no problema de *Named Entity Recognition - NER*. Em Vieira et al. (2015) o CRF é utilizado para analisar comentários sobre produtos e verificar marca e modelo a que se referem. É fornecida uma quantidade pequena de dados de treinamento e a partir destas chamadas sementes anotadas, o modelo CRF é criado, aplicado a sentenças não rotuladas, encontrando novos modelos de produtos, que são adicionados ao conjunto de sementes e re-gerado o modelo CRF, terminando quando não forem mais encontradas sementes.

O proposto no trabalho é o método *Product Model Number Spotter - ModSpot*, o qual gera um modelo CRF sobre uma pequena quantidade de dados e em uma segunda etapa utiliza-se do modelo gerado para descobrir novos produtos. Esses novos produtos são adicionados ao conjunto de sementes para que o CRF seja novamente treinado. Isso ocorre até que não encontre mais novas sementes. Os resultados mostram-se com melhor avaliação utilizando-se desse processo de retreinamento do modelo CRF, quando comparado a utilização de um modelo gerado pelo CRF sem retreinamento.

Para fins de extração de conteúdo de uma página web, pode-se focar nos segmentos relevantes de uma página, distinguindo título, autor, conteúdo, dos comentários, anúncios, ou então a distinção entre imagem, descrição e preço de um produto, por exemplo. Para isso, o CRF foi aplicado como uma ferramenta de segmentação em Fu et al. (2010), em

que são definidas *features* que determinam distâncias entre blocos, características do texto (quantidade de números no texto, pontuação...) e características de layout (tags HTML, como H1, H2, a, li...), ou criando uma variação do modelo CRF, transformando o modelo linear em um modelo de duas dimensões, em Zhu et al. (2005), para realizar essa tarefa.

O Fu et al. (2010) transforma as informações da página web, que estão dispostas em duas dimensões, em informações serializadas. Isso se dá filtrando nodos que contenham informações úteis, seguindo alguns critérios definidos no trabalho, como número mínimo de palavras no nodo, mínimo de urls, etc., e serializando o texto dos nodos mantidos. Feita a serialização, são definidos *labels* e *features* para o treinamento e aplicação do CRF. As labels definidas, ou seja, as informações que o trabalho extrai são: título, autor, conteúdo, comentário, propaganda e outros. Foram criadas *features* que caracterizam o tamanho do texto, características como conter pontuação e dígitos e informações de layout, verificando a existência de determinadas tags HTML.

Os experimentos foram realizados avaliando separadamente a contribuição de cada tipo de *feature* na qualidade da extração: as *features* de layout foram as que demonstraram maior contribuição nos resultados, mas ainda assim, utilizar todas as *features* mencionadas é útil pois aumenta o índice de *f-measure*.

Já no trabalho de Zhu et al. (2005) é mantida a disposição da informação em duas dimensões. Para continuar trabalhando dessa forma, o CRF teve que ser adaptado para o que se propõe no trabalho: 2D CRF. Esse vem a ser um modelo que trabalha com vizinhanças horizontais e verticais entre blocos de HTML, não sendo necessária a serialização, como realizado em Fu et al. (2010). Nos experimentos realizados, foi avaliada a extração de informações, tais como nome, descrição, preço e imagem, referentes a diferentes páginas de produtos em páginas de *e-commerce*. Os resultados desses experimentos, comparados ao resultado da aplicação do CRF linear, mostraram-se significantemente superiores ao outro.

Em Gong e Liu (2012) o modelo CRF é combinado com técnicas de *Tree Edit Distance* aplicado sobre a estrutura DOM, mais uma alternativa para a realização da tarefa de segmentação de páginas e Extração de Informação. Uma vez criada a ferramenta de segmentação, são realizados experimentos para extração de blocos de notícias.

Inicialmente, o CRF é utilizado para segmentação da página entre: bloco de informação, bloco de navegação, bloco de interação, ou nenhum desses. Para isso, é levado em consideração (*features*): a profundidade de cada nodo, ou seja, em que nível da árvore DOM se ele encontra; a TAG HTML daquele nodo; propriedades semânticas, fazendo

uma análise básica do conteúdo para classificá-lo entre título (textos curtos), resumo e conteúdo (textos longos, vídeos, imagens); e tipo de *hyperlink*: sem endereço, com endereço para o próprio site, ou com endereço para outro site. Feito isso, tenta-se localizar blocos repetitivos com o uso de *Tree Edit Distance*. Caso note-se semelhança entre os blocos são agrupados em um só.

Foram realizados experimentos com trinta sites diferentes, contendo dez páginas de cada site, totalizando trezentas páginas. Os resultados mostraram alto *F-measure* na tarefa de segmentação e classificação dos blocos das páginas (bloco de informação, navegação, interação e outros), porém não se executa tarefas de Extração de Informação propriamente dita.

Ainda trabalhando com o conceito da informação estar disposta em duas dimensões, em Pinto et al. (2003), o modelo CRF é aplicado à Extração de Informação a partir de tabelas em páginas Web. Conforme colocado pelos autores, o problema da extração de informação em tabelas pode ser dividido em 6 tarefas: Localizar a tabela, identificar posição das linhas, das colunas, segmentar a tabela em células, rotulas as células como dado ou cabeçalho e associar células a seus cabeçalhos. O artigo trata das duas primeiras tarefas: localização de uma tabela e posicionamento das linhas. Dado um documento, o CRF é utilizado para verificar linha a linha se qual a sua função na tabela e consequentemente localizar o início e fim desta.

Os rótulos definidos para a classificação das linhas do documento são divididos em categorias (Figura 3.2), como a não ocorrência de tabela: não-tabela, linha em branco, separadores; labels que indicam que a linha faz parte de um cabeçalho: título, cabeçalho principal, cabeçalho de tabela, sub cabeçalho e seção de cabeçalho; linha com dados; legendas: nota de rodapé da tabela, legenda de tabela.

Para a execução do CRF, foram definidas *features* relacionadas a existência de dígitos ou caracteres especiais em sequência na linha, palavras-chave, layout, etc. Experimentos realizados sobre páginas web mostram ótimos resultados ao rotular as linhas do documento.

Uma das principais tarefas da Extração de Informação é a detecção de relações entre palavras para, a partir de dados não estruturados, extrair informações estruturadas. A exemplo disso, o *TextRunner* (ETZIONI et al., 2008) se utiliza do CRF para aprender o padrão sintático das sentenças e então extrair tuplas do tipo Entidade-Relação-Entidade, como, por exemplo, *Paris - CapitalOf - France*. Em uma primeira fase, o sistema utiliza CRF para rotular semanticamente a frase sendo analisada, vide Figura 3.3. Após

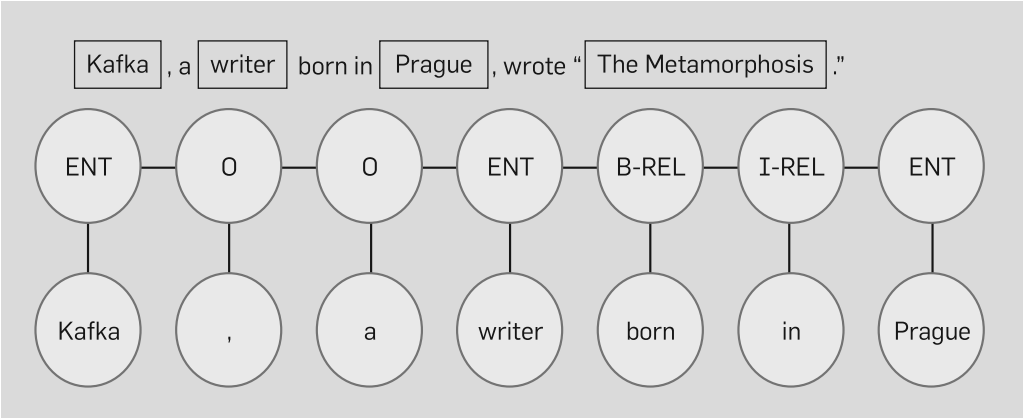
Figura 3.2: Exemplo de tabela com rótulos anotados

```
<TITLE>                Principal Vegetables for Fresh Market:  Area Planted and Harvested
<TITLE>                                by Crop, United States, 1997-99 1/
<TITLE>                                        (Domestic Units)
<SEPARATOR>     ------------------------------------------------------------------------------
<SUPERHEADER>                  :          Area Planted          :        Area Harvested
<SUPERHEADER>         Crop     :--------------------------------------------------------------
<TABLEHEADER>                  :   1997   :   1998   :   1999   :   1997   :   1998   :  1999
<SEPARATOR>     ------------------------------------------------------------------------------
<SUBHEADER>                    :                              Acres
<SEPARATOR>                    :
<DATAROW>       Artichokes 2/  :    9,300      9,700      9,800      9,300      9,700      9,800
<DATAROW>       Asparagus 2/   :   79,530     77,730     79,590     74,030     74,430     75,890
<DATAROW>       Beans, Lima    :    2,700      3,000      3,200      2,500      2,000      2,900
<DATAROW>       Beans, Snap    :   90,260     94,700     98,700     82,660     87,800     90,600
<DATAROW>       Broccoli 2/    :  130,800    134,300    137,400    130,800    134,300    137,300
<SECTIONHEADER> Brussels       :
<SECTIONDATAROW> Sprouts 2/    :    3,200      3,200      3,200      3,200      3,200      3,200
<DATAROW>       Cabbage        :   77,950     79,680     79,570     75,230     76,280     74,850
```

Fonte: Pinto et al. (2003)

isso, o extrator do *TextRunner* forma tuplas com três strings, onde a primeira e a terceira são entidades e a segunda um componente de relação entre as entidade. No exemplo da Figura 3.3, uma tupla extraída seria *Kafka - born in - Prague*.

Figura 3.3: Exemplo de frase anotada pelo CRF. ENT significa marcação de entidade; B-REL, o início de uma relação; e I-REL a continuação da sequência



Fonte: Etzioni et al. (2008)

Após a extração, o sistema indexa as tuplas com o Lucene[5], um indexador e motor de pesquisa de alta performance, permitindo assim a consulta por tuplas contendo alguma entidade em específico (p.ex. Paris), par de entidade (p.ex. Microsoft e IBM) ou relação (p.ex. born). Os experimentos foram realizados sobre uma base de dados com 120 milhões de páginas Web e, para avaliação da precisão, foram selecionadas randomicamente tuplas extraídas, atingindo um índice de setenta e cinco porcento.

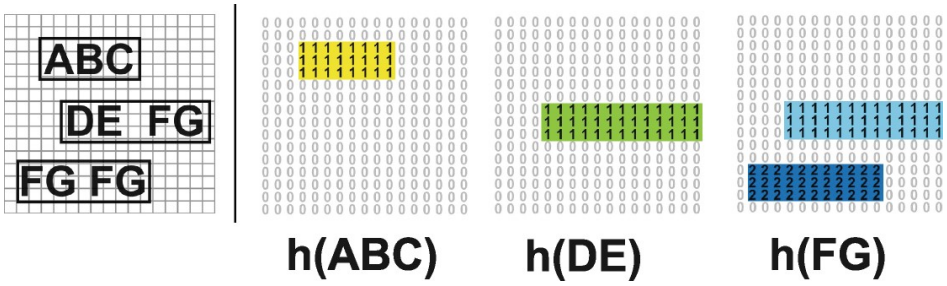
[5]<https://lucene.apache.org/>

3.3 Extração de Informações com outras técnicas

Em se tratando de correlação de informações em páginas, o trabalho Nguyen, Nguyen e Freire (2008) faz a extração de rótulos relacionados a campos de *webforms* no sistema chamado *LabelEx*. A informação contida nesses rótulos é importante para a recuperação de informação na Web oculta, onde é necessário o preenchimento de campos de busca para recuperar documentos. Para a extração dos rótulos, foram testados os classificadores *Naive Bayes*, *Decision Tree*, *Suport Vector Machine - SVM* e *Regressão Logística*. Inicialmente são gerados "mapeamentos candidatos" entre rótulos e campos de entrada próximos uns aos outros. Então é realizada uma poda de algumas relações false, tendo sido utilizado o classificador *Naive Bayes*, que dentre os testados, foi o que obteve melhores resultados para essa tarefa. Em seguida, para a localização da melhor relação entre rótulo e campo de entrada, o algoritmo *Decision Tree* foi escolhido, por obter os melhores resultados. Para fins de experimento, o *LabelEx* é aplicado sobre formulários de diferentes domínios, obtendo altos índices de *F-measure*.

Outra proposta para extração de dados independente de estrutura de sites é feita por Gogar, Hubacek e Sedivy (2016) com a utilização de Redes Neurais Profundas para criação de um modelo combinando informações textuais e de *layout*. Inicialmente a página é renderizada e salvo um *screenshot* e também a árvore DOM. Ao invés de trabalhar com os nodos da árvore DOM ou com a posição desses nodos na página renderizada, o proposto é criar uma grade na qual as palavras são mapeadas individualmente (Figura 3.4).

Figura 3.4: Mapeamento de palavras na página. Figura à esquerda, nodos com as palavras "ABC", "DE"e "FG"; nas demais figuras, o mapeamento das palavras h()



Fonte: Gogar, Hubacek e Sedivy (2016)

Páginas de *e-commerce* são submetidas a esse processo, que serve de entrada para uma rede profunda. Por fim, as regiões da página são classificadas como nome do produto, imagem e preços, possibilitando assim a extração das informações. Nos testes realizados atingiu-se alta acurácia.

Quanto à extração de dados de *Calls-For-Papers* (*CFPs*) enviados por e-mail, Li et al. (2013) trabalha na extração de afiliações, ou seja, organizações às quais pesquisadores estão vinculados. O método proposto combina *Named Entity Recognition* (NER) e informações de *layout* para a realização da extração. O primeiro é utilizado para detectar nome do pesquisador, da organização e outros; informações de *layout* são utilizadas para a diferenciação de áreas com informações relevantes para a extração, das demais regiões. Esse trabalho não faz extração das datas dos CFPs.

3.4 Sumário dos Trabalhos Relacionados

Esse Capítulo descreveu os trabalhos estudados para o desenvolvimento do proposto nesse trabalho e também relacionados à tarefa de extração de dados da web de um modo mais amplo. Na Seção 3.1 é dado enfoque a trabalhos diretamente relacionados a conferências e a extração e disponibilização de suas informações, principalmente referente a datas. Em seguida, na Seção 3.2 é abordada a extração de informações da web em um contexto geral, porém focando na técnica de *Conditional Random Fields*. E por fim, na Seção 3.3, trabalhos diversos com o mesmo objetivo, porém utilizando outras técnicas.

Os trabalhos relacionados na seção 3.1 são basicamente ambientes de repositórios de datas cadastradas manualmente, com exceção do *All Call* que, mediante acesso a e-mails do usuário, realiza a extração de datas. Esse trabalho diferencia-se por ser independente de cadastros manuais, ou de acesso a e-mails do usuário, e realiza a extração dos *deadlines* de conferências de forma automatizada.

Os trabalhos mencionados nas seções anteriores fazem uso do *framework* CRF (seção 3.2) ou de outras técnicas de Aprendizado de Máquina (seção 3.3) em tarefas diversas relacionadas à extração de informação, tais como de *Named Entity Recognition*, segmentação de páginas, etc. Porém nenhum desses trabalhos realizam a localização e extração de datas.

Se tratando especificamente dos trabalhos que utilizam CRF, como estes se aplicam a outros domínios, as *features* projetadas possuem outros objetivos. Nesse sentido, o diferencial do presente trabalho está em trabalhar com a extração de datas, mais especificamente, *deadlines* de conferências.

Dentre todos os trabalhos estudados, os que no geral reportaram melhores resultados em seus experimentos foram os que se utilizam do CRF, tendo sido essa a técnica utilizada para o desenvolvimento desse trabalho.

Na Tabela 3.1 é realizado um comparativo entre os trabalhos, mostrando quais técnicas cada um utiliza, sendo que de modo geral, a maioria dos trabalhos relacionados a área encontrados utiliza-se do CRF.

Tabela 3.1: Técnicas utilizadas para extração de informação

	Manual	Busca Textual	Análise Posicional (Layout)	CRF	Tree Edit Distance	Decision Tree	Redes Neurais Profundas
WikiCFP	X						
Conference Alarm	X						
AllCall (CORREIA et al., 2010)		X					
Mattes (2011)			X				
ModSpot (VIEIRA et al., 2015)				X			
Fu et al. (2010)				X			
Zhu et al. (2005)				X			
Gong e Liu (2012)				X	X		
Pinto et al. (2003)				X			
TextRunner (ETZIONI et al., 2008)				X			
LabelEx (NGUYEN; NGUYEN; FREIRE, 2008)						X	
Gogar, Hubacek e Sedivy (2016)							X
Li et al. (2013)			X				

Fonte: O autor

4 PROCESSO DE EXTRAÇÃO DE INFORMAÇÕES DE CONFERÊNCIAS PARA CONSULTA CENTRALIZADA VIA AMBIENTE WEB

Nesse capítulo é descrito o processo completo para a disponibilização das datas para consulta em um ambiente Web, desde a pesquisa de URLs a partir das conferências listadas na Tabela Qualis, download de conteúdo Web, execução da extração com etapas de pré e pós processamento e disponibilização do conteúdo extraído.

4.1 Visão Geral do Processo

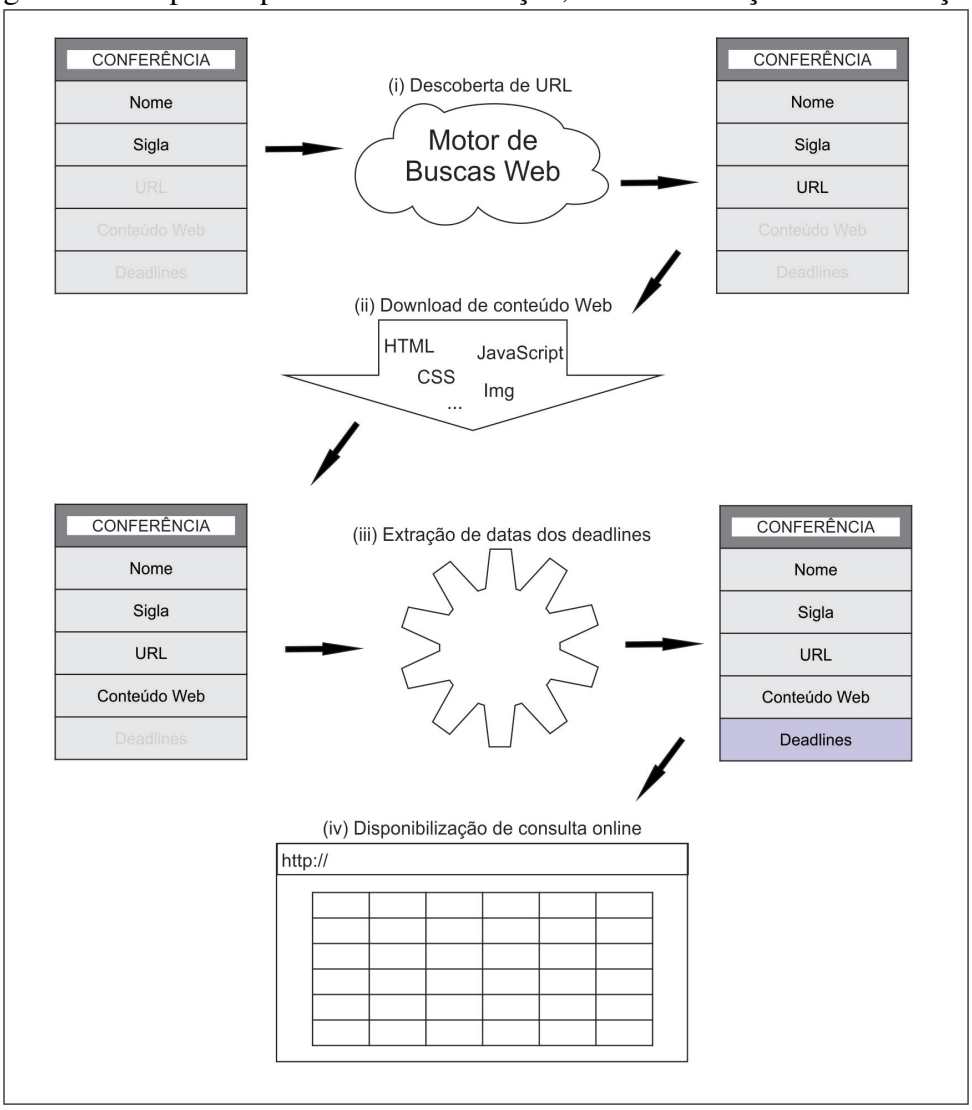
O problema investigado nesse trabalho pode ser resumido como *"Dado um conjunto de conferências de interesse (representadas por seus nomes e siglas) encontre suas páginas web e respectivas datas de interesse"*. A Tabela Qualis (que contém a sigla, o nome e a classificação das conferências) torna-se então o ponto de partida para a resolução desse problema. A fim de atingir o objetivo proposto, definimos um processo composto pelas seguintes etapas apresentadas na Figura 4.1: (i) descoberta das URLs das conferências, (ii) download do conteúdo, (iii) extração das datas, e (iv) disponibilização da informação para consulta online.

A seguir, são definidos termos utilizados no decorrer da dissertação que são necessários ao entendimento da proposta:

- datas de interesse: são os valores das datas limites, ou *deadlines*, para entrega de resumo, do artigo, data de notificação de aceitação, envio da versão final e período da conferência;
- rótulo: trata-se do texto usualmente vinculado a uma data, indicando a qual das datas de interesse aquela data se refere. Um exemplo de rótulo referente ao *deadline* de submissão de resumo é *"abstract submission"*, ou *"abstracts deadline"*.

Para exemplificar, considere a Figura 4.2. A data de interesse "*January 17, 2017*" refere-se ao prazo de submissão do resumo, cujo rótulo é "*Abstracts for full research papers due*" enquanto que a data de submissão do artigo "*January 24, 2017*" está associada ao rótulo "*Full Research papers due*".

Figura 4.1: Etapas do processo de localização, coleta e extração de informações



Fonte: O autor

4.2 Descoberta de URL

Na primeira etapa, a partir de uma lista de conferências de interesse (dada pela Tabela Qualis), é necessário descobrir a URL do site. Isto é feito por meio de consultas a motores de busca na Web. As consultas são compostas pelo ano, sigla e o nome da conferência. As N primeiras URLs retornadas são mantidas para a fase seguinte. Além disso, são excluídas URLs de sites como *facebook*, *wikepedia*, *twitter*, *linkedin*, *github*, etc, que têm baixa probabilidade de ser a página oficial de uma conferência.

Figura 4.2: Exemplo de página de conferência com datas de interesse e rótulos. Extraída da página do EDBT 2017 http://edbticdt2017.unive.it/?important_dates

Fonte: Adaptado de http://edbticdt2017.unive.it/?important_dates

4.3 Download de conteúdo Web

A segunda etapa é responsável pelo *download* do conteúdo das URLs retornadas pela etapa anterior. São baixados recursivamente todos os documentos vinculados na URL informada sob o mesmo domínio, ou seja, URLs que direcionam a outras páginas do mesmo site da conferência, limitando-os a dois níveis. É baixado somente o conteúdo HTML, dessa forma arquivos com extensões mp3, mp4, pdf, ppt, pptx, doc, docx, etc. são ignorados, uma vez que esses não são analisados na etapa de extração e acarretariam um trafego muito maior de rede ao realizar o *download*.

Por exemplo, a URL localizada para uma determinada conferência endereça para uma página HTML. Essa página HTML é baixada e verificadas as URLs (TAGS <a>) que constam nela. Para cada URL que enderece páginas do mesmo domínio, seu conteúdo é baixado e armazenado. Assim se restringe o volume de páginas a serem baixadas e analisadas na próxima etapa.

O interesse em baixar mais de um nível (seguindo as URLs da página inicial) está no fato de que nem sempre as datas de interesse encontram-se na página inicial da conferência, porém nesses casos, há um link para a página com tais informações, geralmente intitulada de *"Important Dates"* ou *"Call for Papers"*.

4.4 Extração de datas de interesse

Nessa terceira etapa, os arquivos baixados na etapa anterior são analisados, as datas são extraídas e armazenadas em uma base de dados. Essa etapa é o foco principal deste trabalho e é detalhada no Capítulo 5.

O principal desafio do trabalho encontra-se nessa etapa, pois além de encontrar as datas em meio ao conteúdo Web, há a necessidade de serem corretamente relacionadas às datas de interesse definidas.

A técnica utilizada para a extração dessas datas foi a *Conditional Random Fields - CRF*. A motivação para a escolha desse *framework* foi sua ampla utilização na literatura para tarefas relacionadas ao processamento de linguagem natural, e principalmente por atender com bons resultados às tarefas de rotulação e segmentação de dados.

4.5 Disponibilização de consulta online

A quarta etapa consiste na disponibilização de um ambiente web para a consulta das conferências e suas datas. Dessa forma, as datas coletadas poderão ser facilmente e gratuitamente acessíveis pela comunidade científica.

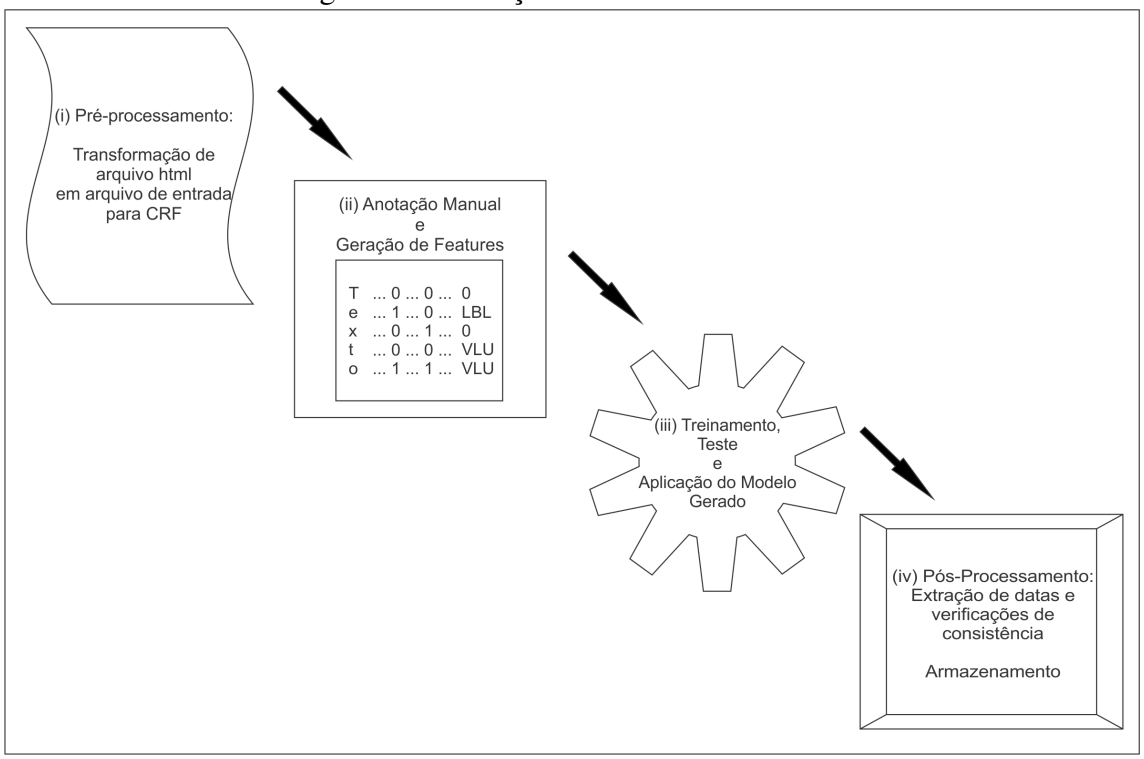
4.6 Sumário

Nesse capítulo é explicado de forma geral o processo definido para a resolução do problema *"Dado um conjunto de conferências de interesse (representadas por seus nomes e siglas) encontre suas páginas web e respectivas datas de interesse"*. Dentre as etapas de localização do site da conferência, download do conteúdo, extração de datas e disponibilização em ambiente centralizado, a etapa que traz as maiores contribuições desse trabalho é a de extração das datas, a qual realiza a difícil tarefa de encontrar as datas e vincular cada rótulo de data de interesse à data correta.

5 EXTRAÇÃO DE DATAS DE CONFERÊNCIAS A PARTIR DE SUAS PÁGINAS WEB

Nesse capítulo, aborda-se com detalhes a etapa de extração das datas (Figura 5.1), passando pelo (i) pré-processamento, onde as páginas são preparadas para a análise; (i) geração das *features* necessárias para o bom funcionamento do algoritmo de rotulação dos dados; (iii) treinamento, teste e aplicação do algoritmo; e (iv) pós-processamento, onde a saída do algoritmo é analisada e da qual são extraídas as datas de interesse.

Figura 5.1: Extração de datas dos *deadlines*



Fonte: O autor

O algoritmo utilizado para a rotulação dos dados em páginas HTML e em seguida extração das datas foi o *Conditional Random Fields - CRF* (LAFFERTY et al., 2001), um *framework* para geração de modelos probabilísticos para segmentação e rotulação de dados. A motivação para a escolha desse *framework* foi sua ampla utilização na literatura para tarefas relacionadas ao processamento de linguagem natural, conforme relatado no Capítulo 3, e principalmente por atender com bons resultados às tarefas de rotulação e segmentação de dados, à qual se propõe.

Ao se trabalhar com essas tarefas, devem ser definidos os possíveis valores de rótulos ou segmentos, também chamados de "Classes". Neste trabalho a tarefa é de rotulação, sendo que as possíveis Classes aplicáveis a cada observação são definidas na Tabela 5.1.

Figura 5.2: Arquivo de treinamento

```
...
dates 0 0 0 0 1 0 0 0 0 0
january 0 0 0 0 0 0 1 0 0 VLU_ABS
14 0 0 0 0 0 1 0 1 0 VLU_ABS
2016 0 0 0 0 0 0 0 0 1 0 VLU_ABS
abstracts 1 0 0 0 0 0 0 0 0 0 LBL_ABS
for 0 0 0 0 0 0 0 0 0 LBL_ABS
full 0 1 0 0 0 0 0 0 0 0 LBL_ABS
research 0 1 0 0 0 0 0 0 0 0 LBL_ABS
papers 0 1 0 0 0 0 0 0 0 0 LBL_ABS
due 1 0 0 1 0 0 0 0 0 0 LBL_ABS
january 0 0 0 0 0 0 1 0 0 VLU_PPR
21 0 0 0 0 0 1 0 1 0 VLU_PPR
2016 0 0 0 0 0 0 0 0 1 0 VLU_PPR
full 0 1 0 0 0 0 0 0 0 0 LBL_PPR
research 0 1 0 0 0 0 0 0 0 0 LBL_PPR
papers 0 1 0 0 0 0 0 0 0 0 LBL_PPR
...
```

Fonte: O autor

Para a execução dessa importante etapa do processo definido nesse trabalho, uma tarefa de pré-processamento (i) se faz necessária (Figura 5.1), pois a ferramenta de CRF utilizada necessita que o arquivo de treinamento esteja em um formato específico: cada linha do arquivo é formada por N colunas. Conforme ilustrado na Figura 5.2, a primeira coluna representa um *token*, ou seja, o texto do HTML é extraído (removendo-se as *tags*) e cada linha desse arquivo de treinamento representa uma palavra do texto extraído; a N-ésima coluna de cada linha representa a Classe manualmente anotada (Tabela 5.1); as colunas intermediárias representam as *features* implementadas e abordadas a seguir.

Como entrada para o treinamento desse algoritmo, foram anotadas as datas de interesse e rótulos. Uma vez gerado o modelo, este é aplicado a todas as conferências. Esse processamento gera um arquivo semelhante ao da Figura 5.2, porém com uma coluna a mais, contendo a predição do CRF, de onde as datas de interesse são obtidas no pós-processamento.

No ConfTracker, o modelo CRF foi treinado para detectar dez possíveis classes de saída, referentes aos rótulos e valores das datas de interesse, conforme relacionado na Tabela 5.1.

Previamente foram definidos *tokens* que podem referenciar os rótulos das datas de interesse (LBL_*). Por exemplo, os *tokens* $abstract$, $paper$, $submission$ são associados aos rótulos do deadline de submissão de resumo; *tokens* $camera$, $ready$, due, associados aos rótulos de confirmação de aceitação, etc.

Para a detecção das classes de rótulo e valores de datas de interesse, é necessá-

Tabela 5.1: Classes definidas para os rótulos de cada data de interesse

Deadline	**Classe de Rótulo**	**Classe de Data**
Resumo	LBL_ABS	VLU_ABS
Artigo	LBL_PPR	VLU_PPR
Notificação de aceitação	LBL_ACC	VLU_ACC
Versão final	LBL_CAM	VLU_CAM
Período da Conferência	LBL_EVE	VLU_EVE

Fonte: O autor

rio definir os atributos (ou *features*) a serem consideradas pelo CRF. Foram criadas nove *features* para a detecção dessas classes, listadas na Tabela 5.2: seis delas para distinguir entre os rótulos das diferentes datas de interesse e as outras três para detectar datas. Cada uma dessas *features* representa uma coluna intermediária no arquivo de treinamento (Figura 5.2),

Tabela 5.2: *Features* implementadas como parte da entrada do algoritmo de CRF

Tipo	*Feature*
Feature de Rótulo	*Token* está associado a Submissão do Resumo?
	Token está associado a Submissão do Artigo?
	Token está associado a Notificação de Aceitação?
	Token está associado a Submissão da Versão Final?
	Token está associado ao Período da Conferência?
	Token é uma Sigla de Conferência?
Feature de Data	*Token* pode ser um Dia?
	Token pode ser um Mês?
	Token pode ser um Ano?

Fonte: O autor

A exemplo do que foi explicado na Seção 2.2, cada uma dessas *features* geram valores $b(x, i)$, que caracterizam os dados que o CRF irá rotular:

$$b(x,i) = \begin{cases} 1 & \text{se a observação na posição } i \text{ está associada a Submissão do Resumo} \\ 0 & \text{caso contrário} \end{cases}$$

Esse exemplo mostra a geração da primeira *feature* listada na Tabela 5.2, sendo que um *token* "está associado" a um determinado *deadline* quando este encontra-se dentre uma lista pré-determinada de tokens para cada *deadline*. Da mesma forma são geradas as demais *features* de Rótulo para cada *token*.

Já as *features* de Data avaliam *tokens* que podem compor uma data válida, ou seja, a *feature* que determina se um *token* pode ser um Dia ativa-se quando o *token* é um

número de 1 a 31; já a *feature* que sinaliza se um *token* é um mês é ativa quando este for um número de 1 a 12, ou caso seja um mês escrito por extenso ou abreviado com no mínimo três letras; e a *feature* de Ano, ativa-se quando o *token* é um número de dois ou quatro dígitos. Estas regras são validadas por meio de expressões regulares.

O método CRF também necessita de um *template*, que permite a criação de uma janela de contexto, ou seja, ao classificar um *token* (ou aprender a fazê-lo) considerar também os x *tokens/features* anteriores e próximos.

Nesse arquivo, exemplificado na Figura 5.3, cada linha representa um template (linhas iniciando com # são comentários) no formato prefixo, id e regra: "U_i : $\%x[row, col]$", onde *U* é o prefixo; *i* é o id; *%x* é fixo; *row* é a linha, que pode ser definida com números negativos (indicando linhas prévias ao *token* sendo classificado), número zero (própria linha) e números positivos (próximas linhas); *col* é a coluna, indicando a *feature* a ser analisada e assumindo valores com a mesma lógica dos valores de *row*.

Figura 5.3: Template

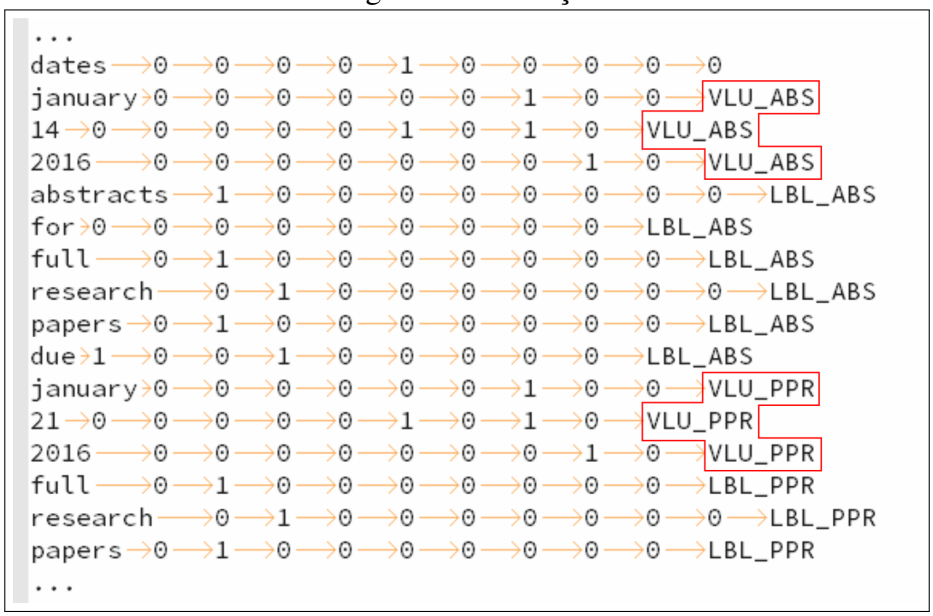
```
...
# token
U49:%x[0,0]
# abstract
U50:%x[0,1]
# papper
U51:%x[0,2]
# acceptance
U52:%x[0,3]
# camera ready
U53:%x[0,4]
# event period
U54:%x[0,5]
# day
U85:%x[0,6]
# month
U86:%x[0,7]
# year
U87:%x[0,8]
# conference initials
U88:%x[0,9]
...
```

Fonte: O autor

Realizada a rotulação pelo CRF, a partir do arquivo de saída fornecido pelo algoritmo é executado o pós-processamento, que verifica as classes atribuídas pelo algoritmo em busca de agrupamentos de Classes do tipo VLU_*. Por exemplo, na Figura 5.4, o CRF marcou os valores "*january*", "*14*" e "*2016*" como VLU_ABS, e "*january*", "*21*" e "*2016*" como VLU_PPR. Ou seja, para o *deadline* de submissão de abstract, a data extraída será "14/01/2016" e para o *deadline* de submissão de artigo, a data "21/01/2016".

Feita a extração, as datas ainda passam por uma verificação de consistência, em

Figura 5.4: Extração



Fonte: O autor

que precisam atender a ordem cronológica dos diferentes prazos. As datas de interesse de uma conferência obedecem a uma ordem *deadline de submissão de resumo* $<$ *deadline de submissão de artigo* $<$ *data de notificação de aceitação* $<$ *deadline de submissão de versão final* $<$ *período da conferência*, logo, as datas extraídas precisam obedecer a essa regra. Caso uma data não atenda esta ordem, ela é desconsiderada e não é armazenada.

5.1 Sumário

Nesse Capítulo, a terceira etapa do processo, envolvendo a descoberta, extração e verificação de consistência das datas é detalhadamente explicada. No pré-processamento os arquivos html são preparados para o CRF, juntamente com as *features* e classes anotadas manualmente (o processo de treinamento e testes é detalhado no Capítulo 6). Realizada a anotação, a rotina de pós-processamento analisa as Classes atribuídas aos *tokens*, extrai as datas e verifica consistência.

6 EXPERIMENTOS

Nesse Capítulo são descritos os experimentos realizados, ferramentas utilizadas, a metodologia de avaliação, configuração dos dados do *Gold Standard* e resultados das avaliações. Os experimentos aqui descritos foram realizados sobre conferências selecionadas randomicamente, a partir das quais foi montado o *Gold Standard* e realizadas avaliações. Para a avaliação da qualidade da extração das datas, foram realizados experimentos comparando as datas extraídas pelo CRF às datas esperadas. Além disso, os resultados do CONFTRACKER foram comparados aos de um *baseline*. Nossos experimentos foram realizados sobre conferências da Ciência da Computação. Aborda-se também, em outro experimento, a avaliação da etapa de busca das URLs de cada conferência. Após detalhar a execução desta etapa do processo, são apresentadas avaliações e resultados alcançados.

6.1 Avaliação do Método de Extração de Datas

Nessa seção é dado enfoque à execução da etapa de extração das datas. Expõe-se detalhadamente o desenvolvimento dos experimentos e avaliações.

6.1.1 Materiais e Métodos

6.1.1.1 Gold Standard

O *Gold Standard* é composto pelos dados gerados por especialista, representando o conjunto de informações corretas e sendo a resposta esperada ao aplicar-se uma técnica de extração de dados. Avaliações da qualidade da extração baseia-se na comparação entre os dados obtidos pela extração e os dados do *Gold Standard*.

A partir da Tabela Qualis[1], foram escolhidas aleatoriamente oitenta conferências, as quais foram utilizadas para análise e suporte ao desenvolvimento e também para realização de avaliações da extração. Sobre as conferências sorteadas, cada uma das datas disponíveis foi pesquisada manualmente para a criação do *Gold Standard*, que contém as datas esperadas. Como o objetivo desse experimento é avaliar apenas a qualidade da extração de datas, as etapas da seleção das URLs das conferências e *download* do con-

[1] <https://www.capes.gov.br/images/stories/download/avaliacao/Comunicado_004_2012_Ciencia_da_Computacao.pdf>

teúdo web, que podem retornar o conteúdo Web incorreto para alguma conferência, foram suprimidas, utilizando-se o conteúdo Web localizado e feito o download de forma monitorada.

6.1.1.2 Ferramentas

O sistema foi desenvolvido na linguagem C# da plataforma .Net, sobre o sistema operacional Windows 10, obtendo-se suporte de algumas ferramentas para tarefas específicas, tais como:

- Wget[2]: para download do conteúdo Web;
- CRFSharp[3]: utilizada para aplicação da técnica de CRF;

6.1.1.3 Métricas de avaliação

As métricas utilizadas para avaliação da qualidade da extração são:

- Precisão, sendo a razão entre a quantidade de datas corretas extraídas (R_q) e o número de datas extraídas (S_q):

$$Precision = \frac{|R_q|}{|S_q|}$$

- Revocação, a razão entre a quantidade de datas corretas extraídas R_q e o número de datas de interesse no *Gold Standard* (R):

$$Recall = \frac{|R_q|}{|R|}$$

- F-measure, a média harmônica entre *Precisão* e *Revocação*:

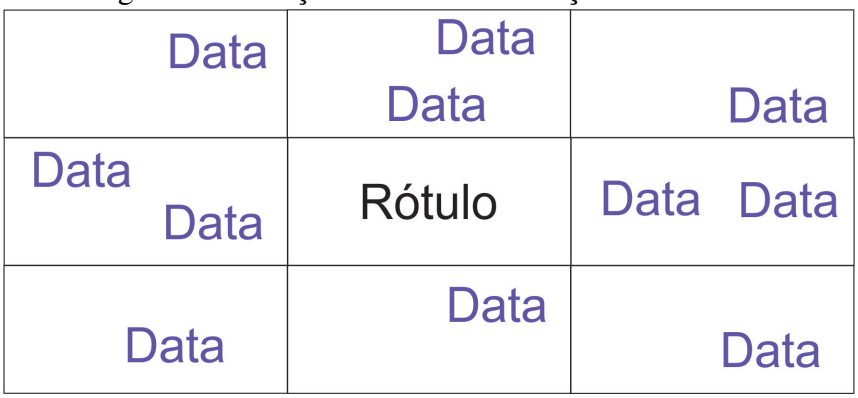
$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

6.1.1.4 Baseline

Os resultados são comparados com os de um *baseline* implementado que baseia-se em Mattes (2011) que propôs uma forma de automatizar o processo de extração de datas que explora a posição dos rótulos e dos valores. A páginas são renderizadas em memória,

[2]<https://www.gnu.org/software/wget/>
[3]<https://github.com/zhongkaifu/CRFSharp>

Figura 6.1: Posições de datas em relação a um rótulo



Fonte: O autor

juntamente com CSS e Javascript, para que as posições [x, y] de rótulos e valores de datas sejam descobertos. Essa técnica, além de avaliar o conteúdo na forma como ele estaria sendo exibido, tenta emular a percepção humana de uma página exibida. São criados pares de "nodos rótulo" e "nodos data", relacionando todos esses nodos encontrados na página e definidos valores de importância para a posição das datas em relação aos rótulos, conforme Figura 6.1. As datas recebem um *score* em função da sua posição em relação ao rótulo selecionado.

6.1.1.5 Procedimento

Uma vez extraídas as datas, é executado o módulo de avaliação que faz os cálculos de precisão, revocação e *f-measure*, medindo, assim, o desempenho da extração dos métodos ConfTracker e *baseline*. Além disso, avaliamos os resultados da combinação do ConfTracker com o *baseline*. Essa combinação foi obtida a partir dos resultados do ConfTracker, preenchendo-se as lacunas sem datas com as datas encontradas pelo *baseline*. Essa estratégia foi escolhida, pois conforme resultados apresentados e discutidos na Seção 6.1.2, o ConfTracker foi bastante superior ao *baseline*, tendo-se assim a extração do ConfTracker como resultado principal, sendo complementado pelo *baseline*.

6.1.2 Resultados e Discussão

Nessa seção são expostos os resultados de todos os experimentos realizados, bem como comparativos entre eles e comentários sobre essas informações.

A Tabela 6.1 mostra os resultados das execuções do *baseline*, do ConfTracker

e da combinação entre eles, aplicadas sobre as oitenta conferências utilizadas. Considerando-se a média das métricas para todas as datas de interesse, O CONFTRACKER obteve a melhor precisão (0,804) e o melhor *F-measure* (0,703). A precisão atingida pelo CONFTRACKER foi 54,8% maior que a do *baseline*, tendo atingido 0,80 contrastando com o índice de 0,52 do *baseline*. A Revocação resultante do CONFTRACKER teve um percentual de superioridade próximo ao da Precisão: 56,2%, tendo o CONFTRACKER atingido 0,63 enquanto o *baseline* obteve 0,52. Comparando-se os índices de *F-measure* da extração executada pelo *baseline* aos da extração do CONFTRACKER, nota-se uma superioridade estatisticamente significativa do CONFTRACKER. Isso é comprovado pelo Teste-T que resultou p-valor de $3{,}94 \times 10^{-8}$.

Comparando o método CONFTRACKER com a combinação CONFTRACKER + *baseline*, observa-se um *F-measure* bastante próximo. Conforme o esperado, a melhor revocação foi obtida pela combinação de métodos CONFTRACKER e *baseline* (0,694), pois dessa forma mais datas de interesse foram localizadas. Contudo, algumas datas localizadas pelo *baseline* não estavam corretas, o que teve um impacto negativo sobre a precisão e, por consequência, sobre o *F-measure*. Assumindo que a precisão seja um índice mais importante que a revocação na resolução do problema definido nesse trabalho, pois vem a ser preferível que uma data não tenha sido extraída do que ter sido extraída erroneamente, a melhor alternativa das três abordagens aqui expostas é a utilização do método CONFTRACKER isoladamente.

Tabela 6.1: Resultados dos experimentos

	Abstract	**Paper**	**Notification**	**Camera Ready**	**Conf Start**	**Conf End**	**Médias**
Baseline							
Precisão	0,296	0,351	0,440	0,448	0,775	0,804	0,519
Revocação	0,242	0,260	0,376	0,400	0,584	0,569	0,405
F-measure	0,266	0,299	0,406	0,422	0,666	0,666	0,454
CRF							
Precisão	**0,823**	**0,620**	**0,800**	**0,784**	**0,898**	**0,898**	**0,804**
Revocação	0,424	0,493	0,637	0,615	0,815	0,815	0,633
F-measure	0,560	**0,549**	**0,709**	**0,689**	0,854	0,854	**0,703**
CRF + Baseline							
Precisão	0,620	0,567	0,712	0,651	0,863	0,863	0,713
Revocação	**0,545**	**0,520**	**0,681**	**0,661**	**0,876**	**0,876**	**0,693**
F-measure	**0,580**	0,542	0,696	0,656	**0,870**	**0,870**	0,702

Fonte: O autor

Avaliando a qualidade da extração por tipo de data de interesse, percebe-se que as datas de início e fim do evento foram extraídas com maior taxa de acerto pelo CONFTRACKER. Isso ocorreu por se tratar de um intervalo de datas, distinguindo das demais *deadlines*, que são datas simples. Essa distinção não impactou os resultados do *baseline*.

A *deadline* de submissão do artigo foi a que teve o pior resultado na extração do CONFTRACKER. Percebeu-se que isso ocorre devido aos termos dos rótulos relacionados a essa data de interesse serem muito genéricos e estarem presentes em muitas outras partes da página, como em títulos e em meio ao texto em geral. Isso dificultou que o algoritmo CRF aprendesse a extrair essa data específica. A mesma tendência se observa com a extração realizada pelo *baseline*.

Limitações. Uma situação que prejudica os resultados aqui apresentados são aquelas em que as datas são exibidas em imagens. Nesses casos não está sendo possível extrair as informações necessárias, pois optou-se em não utilizar técnicas de *Optical Character Recognition* (OCR).

Apesar de serem previstas centenas de padrões de datas, ainda há algumas que não são contempladas, como casos em que consta o dia da semana em meio a data. Por exemplo, *04 (tue), october 2016*. Caso esse padrão também fosse contemplado, a expressão regular poderia se tornar muito genérica, tendo maiores chances de aceitar padrões que não correspondessem a datas válidas. O mesmo vem a acontecer com a detecção de rótulos de interesse. Como as informações em páginas Web são expressas em linguagem natural, há ainda formas de expor o mesmo significado em rótulos diferentes, ainda não previstos nos padrões definidos no sistema.

Existe ainda a questão de datas serem atualizadas nos sites das conferências ao passar do tempo. Dessa forma, faz-se necessário definir uma recorrência periódica para que as etapas (ii) e (iii) da Figura 4.1 sejam re-executadas. Uma estratégia seria executá-las a cada quinze dias para todas as conferências e, ao se aproximar dos *deadlines* de *abstract submission* e *full paper submission* de uma conferência, executar estas mesmas etapas diariamente de forma isolada para aquela conferência. Assim, quando o usuário final consulta o ambiente web, a informação estará atualizada na base de dados.

6.2 Avaliação da Técnica de Descoberta da URL da Conferência

Nessa seção é detalhada a execução da busca pelas páginas das conferências abordando os procedimentos adotados nos experimentos e avaliações.

6.2.1 Materiais e Métodos

Para a realização da avaliação das URLs pesquisadas de forma automática, foi agregada ao *Gold Standard* a URL correta de cada conferência. Desta forma, as avaliações são executadas comparando as URLs pesquisadas pelo sistema àquelas pesquisadas de forma manual.

Para cada conferência, são pesquisadas e armazenadas três URLs, sendo que a primeira delas supostamente seria a URL correta e as demais seriam utilizadas no caso de as datas não serem encontradas no site da primeira URL. A *query* de entrada para o motor de busca é composta pelo ano do qual se buscam as datas, seguido pela sigla e o nome completo da conferência. Por exemplo, para a conferência "*International Conference on Machine Learning*", é montada a *query* "*2016 ICML International Conference on Machine Learning*", sendo "*2016*" o ano, "*ICML*" a sigla, e "*International Conference on Machine Learning*" o nome da conferência. O motor de busca utilizado foi o *Google*, que disponibiliza uma *API* para pesquisas, facilmente utilizável na linguagem C#.

Então, para a avaliação, é comparada a URL de cada uma dessas três posições com a URL do *Gold Standard*, para se obter a acurácia das URLs retornadas em cada posição, bem como uma acurácia geral, ou seja, a verificação de a URL do *Gold Standard* estar dentre alguma das três posições.

6.2.2 Resultados e Discussão

Nessa seção, são expostos os resultados das buscas de URLs, de acordo com o descrito a cima, sendo que para uma mesma conferência, mais de uma posição pode conter uma URL correta. Nenhuma URL retornada se repete dentre as três posições, porém mais de uma delas pode estar direcionada ao mesmo domínio. Por exemplo, acontece o seguinte dentre as URLs retornadas para a Conferência *ICML - International Conference on Machine Learning*:

- <http://icml.cc/>
- <http://icml.cc/2016/?page_id=1367>

Na Tabela 6.2, são exibidos os resultados da busca de URLs, sendo que na primeira posição foram retornadas 58 URLs corretas, das 80 Conferências, resultando em uma precisão de 72,5%. Já dentre as três URLs retornadas para cada Conferência, foram

encontradas corretamente as URLs de 65 (81,25% de precisão). E para 15 Conferências, a URL não foi retornada dentre as três primeiras posições da consulta (18,75%).

Tabela 6.2: Resultados da extração de URLs

	Número de URLs Corretas	**Percentual em relação as 80 Conferências**
Posição 1	58	72,5%
Posição 2	11	13,75%
Posição 3	17	21,25%
URL retornada corretamente em ao menos uma das posições	65	81,25%
Não encontradas corretamente	15	18,75%

Fonte: O autor

Nota-se que para a grande maioria das conferências para as quais foi retornada ao menos uma URL correta (65), essa já se encontrava na primeira posição (58 = 89,2%). É importante salientar que ocorrendo o acerto na primeira posição, dificilmente a informação estará correta nas posições seguintes, visto que o motor de busca não repete resultados. Porém existem raros casos em que também é possível que a URL esteja correta em mais de uma posição, em que o motor de busca esteja retornando mais de uma URL para o mesmo site, porém páginas diferentes.

Nos casos em que não são encontradas URLs corretas em nenhuma das três posições ocorreu de serem retornados links para eventos de outros anos, ou ainda para outros sites que remetam parcialmente aos mesmos termos pesquisados para a conferência.

6.3 Sumário

Nesse Capítulo são descritos os experimentos realizados na extração de datas e coleta de URLs, detalhando ferramentas, metodologias de avaliação e resultados das avaliações. Observam-se bons resultados nas duas tarefas, etapas do processo definido para o CONFTRACKER.

Na busca pelos sites das conferências enfrenta-se o problema de nem sempre o motor de buscas recuperar as URLs corretas, ainda assim, encontra-se mais de oitenta porcento das conferências dentro dos três primeiros resultados da pesquisa.

Quanto à extração de datas das páginas já baixadas, verifica-se a comparação da extração realizada de datas pelo CONFTRACKER comparada a implementação inspirada no *baseline*, que mostra uma melhora significativa nos resultados obtidos, atingindo uma

precisão média de 0,80 enquanto o *baseline* atinge 0,52.

Limitações foram encontradas quando o site da conferência apresenta as informações em imagens, não sendo também detectadas datas e rótulos em formatos ou termos não previstos.

7 CONCLUSÃO

Para o desenvolvimento do trabalho aqui exposto, foram estudados trabalhos relacionados à coleta de dados em páginas web, extração de datas e utilização do *fremework* CRF. A partir do estudo, foi estabelecido um processo com suas etapas bem definidas, dando enfoque à etapa de Extração de Informação das páginas web. Para avaliar o método de extração de datas, algumas conferências foram selecionadas randomicamente a partir da Tabela Qualis, sobre as quais foi gerado o *Gold Standard*. Uma técnica de *baseline*, a utilização de CRF e a combinação entre ambas abordagens foram implementadas para realização de avaliação de qualidade da extração, percebendo-se uma diferença significante entre o *baseline* e a abordagem proposta em CONFTRACKER, sendo essa última a melhor das alternativas avaliadas.

Utilizando-se dessa melhor abordagem, o processo foi aplicado a todas as conferências da Tabela Qualis, sendo que as datas extraídas foram disponibilizadas para consulta pela comunidade por meio do endereço <http://inf.ufrgs.br/conftracker>.

Como exposto, ambas as técnicas de extração utilizadas para relacionar rótulos e datas depende imensamente da capacidade de localizar as informações em linguagem natural (rótulos) e não estruturada utilizada nas páginas Web. Trabalhos futuros podem ser realizados no sentido de (i) aperfeiçoar a geração das *features* utilizadas pelo CRF. Por exemplo, as que definem se um *token* está relacionado a datas de interesse. Hoje os possíveis rótulos de cada data de interesse estão armazenados em uma base de dados, rótulos estes provenientes de conhecimento especialista, e ao executar a função geradora da *feature*, verifica-se se o *token* em questão encontra-se dentre os *tokens* de interesse de cada data de interesse cadastrados na base de dados. Seria possível trabalhar a aplicação de aprendizado de máquina para encontrar os *tokens* relacionados a cada uma dessas data para que o valor gerado pela *feature* seja mais confiável.

Outra sugestão de trabalho futuro, seria (ii) o aperfeiçoamento da fase de descoberta de URLs. No estado atual, uma *query* é submetida ao motor de buscas e, aplicando-se alguns filtros, assume-se que os primeiros itens retornados pelo motor são os resultados corretos. Apesar da execução da tarefa de extração das datas ser a de maior impacto nos resultados, obter as páginas corretas para a extração é um requisito fundamental para os resultados serem confiáveis e até mesmo possíveis de serem extraídos.

Diversas conferências possuem mais de uma *track* de submissão, tais como *Long paper*, *Short paper*, as quais possuem *deadlines* distintos. O presente trabalho busca

extrair os *deadlines* da *track* de *Long paper*, porém sugere-se como trabalho futuro a implementação da distinção e extração dos *deadlines* das diferentes *tracks*.

Ainda que os *deadlines* das conferências sejam as informações principais a serem extraídas dos sites das conferências, outra funcionalidade que pode auxiliar muito o pesquisador, usuário do ambiente de consultas de conferências, é a opção de filtrá-las por tópicos de interesse. Para que isso seja possível, uma possibilidade de trabalho futuro seria (iii) a implementação da extração dos tópicos de interesse divulgados dos sites das conferências, informação a ser agregada à base de dados do CONFTRACKER. Como os tópicos divulgados nas conferências podem ser descritos de diversas formas, mesmo referindo-se a um mesmo assunto, seria indicado o agrupamento de tópicos com diferentes descrições em apenas um item que remeta ao assunto de cada conjunto de tópicos.

Mais diretamente relacionado ao ambiente de consultas acessível pelo usuário, uma sugestão de trabalho futuro é (iv) a implementação de uma representação gráfica das datas das conferências listadas após a filtragem realizada. A organização de uma linha do tempo, por exemplo, onde conste todas as datas das conferências de interesse do usuário, poderia facilitar a visualização dos prazos que ainda lhe restam para a submissão de seu trabalho.

Sobre esse trabalho foi desenvolvido o artigo intitulado *Extração de dados de conferências a partir da Web* (GARCIA; MOREIRA, 2017), o qual foi submetido e aceito no 32º Simpósio Brasileiro de Banco de Dados - SBBD, evento no qual o artigo será apresentado para posterior publicação.

REFERÊNCIAS

CORREIA, F. L. et al. Allcall: An automated call for paper information extractor. In: **Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on**. [S.l.: s.n.], 2010. p. 1–4.

ETZIONI, O. et al. Open information extraction from the web. **Communications of the ACM**, v. 51, n. 12, p. 68–74, 2008.

FINK, G. A. **Markov models for pattern recognition: from theory to applications**. [S.l.]: Springer Science & Business Media, 2014.

FU, L. et al. Conditional random fields model for web content extraction. In: **Computing in the Global Information Technology (ICCGI)**. [S.l.: s.n.], 2010. p. 30–34.

GARCIA, C. A.; MOREIRA, V. P. Extração de dados de conferências a partir da web. In: **Simpósio Brasileiro de Banco de Dados**. Uberlândia, MG: [s.n.], 2017.

GOGAR, T.; HUBACEK, O.; SEDIVY, J. Deep neural networks for web page information extraction. In: ____. **Artificial Intelligence Applications and Innovations: International Conference and Workshops, AIAI**. [S.l.: s.n.], 2016. p. 154–163.

GONG, Y.; LIU, Q. Automatic web page segmentation and information extraction using conditional random fields. In: **Computer Supported Cooperative Work in Design (CSCWD)**. [S.l.: s.n.], 2012. p. 334–340.

LAENDER, A. H. F. et al. A brief survey of web data extraction tools. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 31, n. 2, p. 84–93, jun. 2002. ISSN 0163-5808.

LAFFERTY, J. et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **Proceedings of the international conference on machine learning, ICML**. [S.l.: s.n.], 2001. v. 1, p. 282–289.

LI, X. et al. Automatic affiliation extraction from calls-for-papers. In: **Proceedings of the Workshop on Automated Knowledge Base Construction**. [S.l.: s.n.], 2013. (AKBC '13), p. 97–102. ISBN 978-1-4503-2411-3.

MATTES, J. **Automated Meta-Data Extraction for Confsearch**. Zurich, 2011. Semester Thesis, Swiss Federal Institute of Technology Zurich.

NGUYEN, H.; NGUYEN, T.; FREIRE, J. Learning to extract form labels. **Proceedings of the VLDB Endowment**, v. 1, n. 1, p. 684–694, 2008.

PINTO, D. et al. Table extraction using conditional random fields. In: **Proceedings of the annual international ACM SIGIR conference on Research and development in informaion retrieval**. [S.l.: s.n.], 2003. p. 235–242.

RABINER, L.; JUANG, B. An introduction to hidden markov models. **IEEE ASSP Magazine**, v. 3, n. 1, p. 4–16, Jan 1986. ISSN 0740-7467.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, p. 257–286, Feb 1989. ISSN 0018-9219.

SARAWAGI, S. et al. Information extraction. **Foundations and Trends in Databases**, Now Publishers, Inc., v. 1, n. 3, p. 261–377, 2008.

SOUZA, E. P. d.; PAULA, M. C. d. S. Qualis: a base de qualificação dos periódicos científicos utilizada na avaliação capes. **InfoCAPES Boletim Informativo**, v. 10, n. 2, 2002.

VIEIRA, H. S. et al. A self-training crf method for recognizing product model mentions in web forums. In: **European Conference on Information Retrieval**. [S.l.: s.n.], 2015. p. 257–264.

WALLACH, H. M. Conditional random fields: An introduction. 2004.

ZHU, J. et al. 2d conditional random fields for web information extraction. In: **Proceedings of the International Conference on Machine Learning**. [S.l.: s.n.], 2005. p. 1044–1051.