Extração de dados de conferências a partir da Web

Cássio Alan Garcia, Viviane P. Moreira

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

Abstract. *Choosing the most suitable conference to submit a paper is a task that depends on a number of factors including:* (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; and* (iii) *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 paper presents a method developed to retrieve and extract data from conferences web sites. Our method combines the identification of conference URL and deadline extraction. The retrieved data is stored in a database to be searched with an online tool. The paper also reports on experiments that evaluate the quality of the extracted data, focusing on the deadlines.*

Resumo. *A escolha da conferência adequada para o envio de um artigo é uma tarefa que depende de vários fatores incluindo:* (i) *o tema do artigo 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; e* (iii) *a qualidade da conferência. 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, esse artigo apresenta um método desenvolvido para a coleta e extração de dados de sites de conferências. Este método combina a identificação de URLs de conferências da Tabela Qualis à identificação de deadlines. Os dados coletados populam uma base de dados que poderá ser consultada através de uma ferramenta online. O artigo também relata experimentos que avaliam a qualidade dos dados extraídos, enfatizando a extração dos deadlines.*

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 toma bastante tempo dos pesquisadores.

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) é necessário 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 and Paula, 2002] é um instrumento avaliativo cujo resultado é uma tabela que atribui um grau a conferências e periódicos visando 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 baseia-se na inclusão manual de informações e não contempla o cruzamento de dados com o Qualis. O objetivo deste artigo é preencher essas lacunas, permitindo que a comunidade possa consultar dados atualizados sobre CFPs.

Este artigo apresenta um método automático de coleta e extração de dados de conferências. Os dados extraídos são armazenados em um banco de dados de eventos que permite a busca de acordo com seu tópico, prazos de submissão e Qualis. O foco principal deste artigo é a etapa de extração das datas de interesse a partir do web site da conferência (todas as etapas são detalhadas na Seção 4).

Como cada conferência possui seu próprio website, com diferentes layouts, formatos de datas e rótulos para representar um dado *deadline*, o desafio está em conseguir contornar esta grande variação na forma de expor a informação a fim de extrair corretamente as datas de cada *deadline* do evento. Além de reconhecer estes diferentes padrões de datas e rótulos, é necessário vincular corretamente as datas encontradas a seus rótulos, de forma que a data do *deadline* de submissão de artigo não seja atribuída à *deadline* de submissão de resumo, por exemplo.

2. Trabalhos Relacionados

Nesta Seção abordaremos trabalhos relacionados à extração de datas de interesse de conferências.

2.1. Extração de datas de conferências

Em nossas pesquisas, verificamos a existência de trabalhos que centralizam informações referentes a conferências. Por exemplo, o WikiCFP[3] 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).

[1] `http://www.confsearch.org/confsearch/`
[2] `http://wikicfp.com/cfp/`
[3] `http://wikicfp.com/cfp/`

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.

Já o ConfSearch[4] baseia-se em dados do DBLP[5], um importante repositório de dados bibliográficos da Ciência da Computação. Dados de conferências que não estão disponíveis no DBLP, como os prazos de submissão, precisam ser cadastrados pelos usuários do site.

O presente trabalho difere do WikiCFP 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. Além disso, nenhum dos sistemas pesquisados leva em conta a classificação Qualis dos eventos – sendo esta classificação de grande importância no Brasil.

2.2. Extração de Informações com *Conditional Random Fields*

Conditional Random Fields (CRF) [Lafferty et al., 2001] é um modelo estocástico utilizado habitualmente para etiquetar e segmentar sequências de dados ou extrair informações de documentos textuais. Em Vieira et al. [2015] o CRF é utilizado para analisar comentários sobre de 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.

Para fins de extração de conteúdo de uma página web, pode-se dar ênfase aos 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 um *framework* de segmentação em Fu et al. [2010] e em Zhu et al. [2005]. Foram 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...).

Em Gong and Liu [2012] o *framework* CRF é combinado com técnicas de *Tree Edit Distance* aplicadas sobre a estrutura DOM como 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.

Ainda trabalhando com o conceito da informação estar disposta em duas dimensões, em Pinto et al. [2003], o *framework* CRF é aplicado à extração de informações a partir de tabelas em páginas Web. Com base nas *features* propostas, a tabela é detectada em meio à página em uma primeira etapa e, em seguida, são rotulados títulos, cabeçalhos, linhas contendo dados e suas colunas.

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.

[4] `http://www.confsearch.org/confsearch/`
[5] `http://dblp.uni-trier.de/`

A exemplo disso, o *TextRunner* [Etzioni et al., 2008] utiliza-se do CRF para aprender o padrão sintático das sentenças e então extrair tuplas do tipo Entidade-Relação-Entidade.

2.3. Extração de Informações com outras técnicas

Em se tratando de correlação de informações em páginas, o trabalho Nguyen et al. [2008] faz a extração de rótulos relacionados a campos de *webforms* no sistema chamado *LabelEx*. A informação contida nestes rótulos é importante para a recuperação de informação na Web oculta que requer o preenchimento de formulários recuperar dados de bancos de dados online. Para a extração dos rótulos, foram testados os classificadores *Naive Bayes*, *Decision Tree*, *Suport Vector Machines - SVM* e *Regressão Logística*. Inicialmente são gerados "mapeamentos candidatos" entre rótulos e campos de entrada próximos uns aos outros. O classificador *Naive Bayes* foi o que obteve melhores resultados sendo utilizado para uma tarefa de poda de algumas relações falsas. 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 dentre os classificadores testados. 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 et al. [2016] com a utilização de Redes Neurais Profundas para criação de um modelo combinando informações textuais e de *layout*. Nos experimentos, o modelo é aplicado à extração de informações de produtos no *e-commerce*, tais como nome do produto, imagem e preços, atingindo 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. Este trabalho não faz extração das datas dos CFPs.

3. Definição do Problema e Visão Geral da Solução

O problema investigado neste artigo 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 deste problema. A fim de atingir o objetivo proposto, definimos um processo composto pelas seguintes etapas apresentadas na Figura 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, definiremos os termos utilizado no decorrer do artigo 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 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*".

Na primeira etapa, a partir de uma lista de conferências de interesse (dada pela tabela Qualis), é necessário descobrir a URL do site da conferência. 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. Apenas as três primeiras URLs retornadas são mantidas na 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.

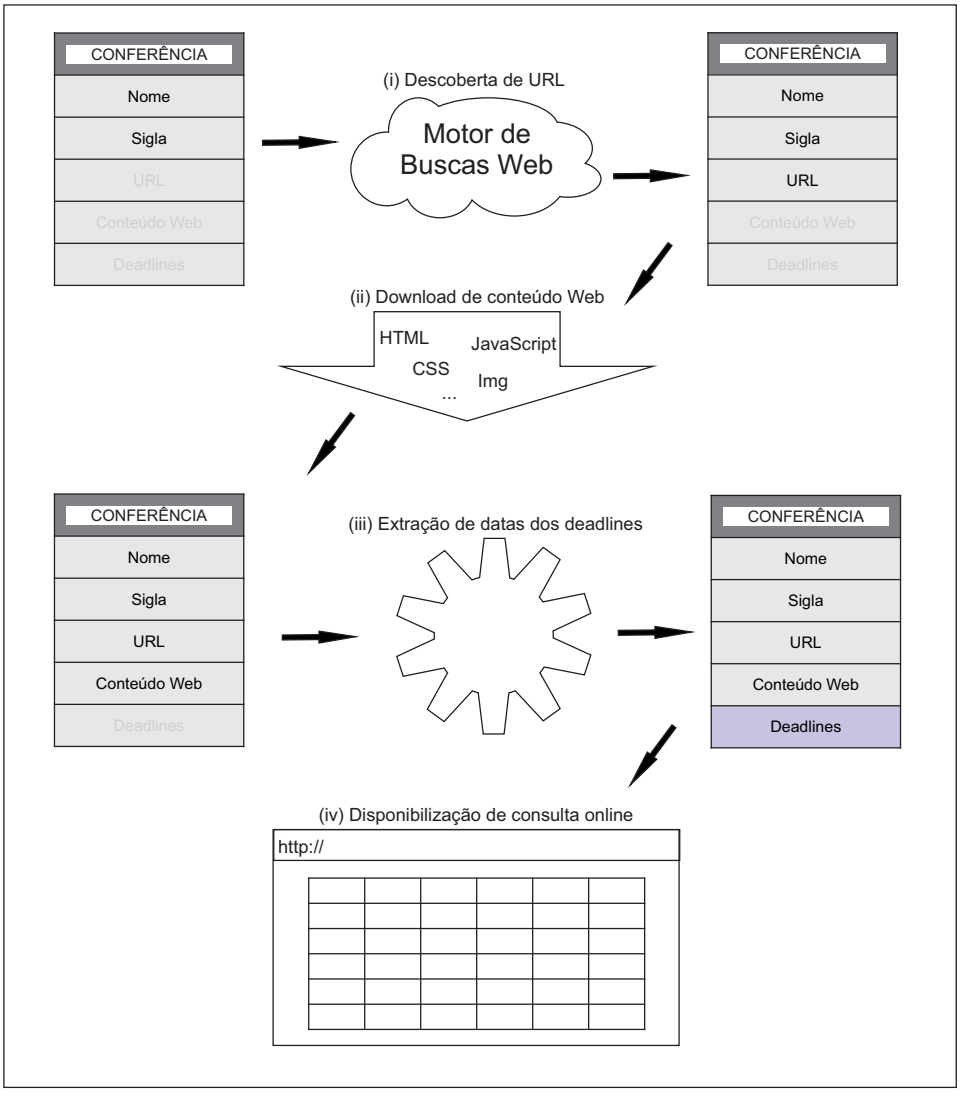

Figura 1. Etapas do processo de extração de datas

Figura 2. Exemplo de página de conferência com datas de interesse e rótulos. Extraída da página do EDBT 2017 http://edbticdt2017.unive.it/?important_dates

Na segunda etapa, é feito o *download* do conteúdo das URLs retornadas pela etapa anterior. São baixados recursivamente todos os documentos vinculados à URL informada, limitando-se a dois níveis. É baixado somente o conteúdo HTML, desta forma, arquivos com extensões mp3, mp4, pdf, ppt, pptx, doc, docx, etc. são ignorados.

Na terceira etapa, os arquivos são analisados, as datas são extraídas e armazenadas. Esta etapa é o foco principal deste trabalho e é detalhada na Seção 4.

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

4. Extração de datas

Para a localização das datas em páginas HTML (Figura 3), utilizamos *Conditional Random Fields* (CRF) [Lafferty et al., 2001], um *framework* para geração de modelos probabilísticos para segmentação e rotulação de dados. A motivação para esta escolha foram os bons resultados dessa técnica aplicados a tarefas similares de Extração de Informações, conforme relatado na Seção 2.

Uma etapa de pré-processamento (i) é necessária (Figura 3), 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. A primeira coluna representa o *token*, ou seja, o texto do HTML é extraído (removendo-se as *tags*) e cada linha deste arquivo de treinamento representa uma palavra do texto extraído; a N-ésima coluna de cada linha representa a Classe manualmente anotada; as colunas intermediárias representam as *features* implementadas e descritas a seguir.

Como entrada para o treinamento do CRF, foram anotadas as datas de interesse e os rótulos. Uma vez gerado o modelo, este é aplicado a todas as conferências. Este processamento gera um arquivo semelhante ao da Figura 4, porém com uma coluna a mais, contendo a predição do CRF. A partir das previsões, as datas de interesse são extraídas.

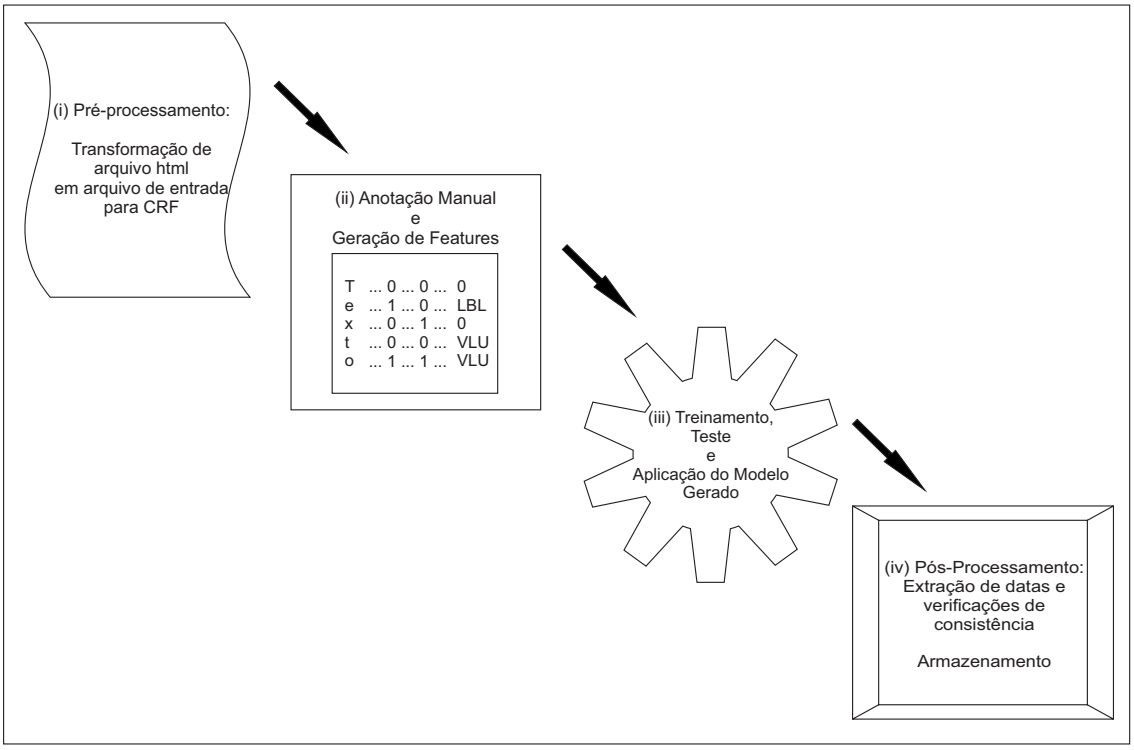

Figura 3. Extração de datas de interesse

No CONFTRACKER, o modelo CRF foi treinado para detectar dez possíveis classes de saída, referentes aos rótulos e valores das datas de interesse:

- deadline de resumo (`LBL_ABS` e `VLU_ABS`),
- deadline de artigo (`LBL_PPR` e `VLU_PPR`),
- notificação de aceitação (`LBL_ACC` e `VLU_ACC`),
- deadline para submissão da versão final (`LBL_CAM` e `VLU_CAM`) e
- período da conferência (`LBL_EVE` e `VLU_EVE`).

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 necessário definir os atributos (ou *features*) a serem considerados pelo CRF. Foram criadas nove *features* para a detecção destas classes: seis delas para distinguir entre os rótulos das diferentes datas de interesse e as outras três para detectar datas. São elas:

- *token* avaliado pertence aos *tokens* previamente definidos para cada uma das cinco datas de interesse; *token* é uma sigla de evento (obtida pela tabela Qualis);
- *token* avaliado é potencialmente um dia, mês ou ano.

O *framework* 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. No arquivo de *templates*, exemplificado na Figura 5, 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.

A partir do arquivo de saída do CRF é executado o pós-processamento, que verifica as classes atribuídas pelo algoritmo, extraindo as datas e fazendo uma verificação de consistência. 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 esta regra.

```
...
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 LBL_ABS
research 0 1 0 0 0 0 0 0 0 LBL_ABS
papers 0 1 0 0 0 0 0 0 0 LBL_ABS
due 1 0 0 1 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 LBL_PPR
research 0 1 0 0 0 0 0 0 0 LBL_PPR
papers 0 1 0 0 0 0 0 0 0 LBL_PPR
...
```

Figura 4. Arquivo de treinamento

```
...
# 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]
...
```

Figura 5. Arquivo de *template*

5. Avaliação Experimental

Para a avaliação da qualidade da extração, foram realizados experimentos comparando as datas extraídas pelo CRF às datas esperadas. Além disso, os resultados do CONFTRACKER foram comparados aos de um *baseline*. Embora o método proposto possa ser aplicado a conferências de diferentes àreas de interesse, nossos experimentos foram realizados sobre conferências da Ciência da Computação.

5.1. Materiais e Métodos

Gold Standard. A partir das conferências da Tabela Qualis[6], foram escolhidas aleatoriamente oitenta, as quais foram utilizadas para a avaliação 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 do experimento é avaliar apenas a qualidade da extração de datas, as URLs das conferências foram selecionadas manualmente.

Ferramentas. O sistema foi desenvolvido na linguagem C# da plataforma .Net, sobre o sistema operacional Windows 10, fazendo uso de algumas ferramentas para tarefas específicas, tais como:

- Wget[7]: para download do conteúdo Web;
- CRFSharp[8]: utilizada para aplicação do *framework* CRF;

Métricas de avaliação. As métricas utilizadas para a avaliação da qualidade da extração são Precisão (P), Revocação (R) e F-mesure (F). Definidas por:

$$P = \frac{|R_q|}{|S_q|}, \quad R = \frac{|R_q|}{|R|}, \quad F = \frac{2 \times P \times R}{P + R}$$

onde R_q é o número de datas corretas extraídas, $|S_q|$ é o número de datas extraídas pelo método sendo avaliado e $|R|$ é o número de datas do *gold standard*.

Baseline. Estes resultados são comparados com os de um *baseline* implementado por nós baseando-se em Mattes [2011] que propôs uma forma de automatizar o processo de extração de datas do ConfSerach explorando a posição dos rótulos e dos valores. A páginas são renderizadas em memória, juntamente com CSS e Javascript, para que as posições [x, y] de rótulos e valores de datas sejam descobertos. Esta 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 ilustrado na Figura 6. As datas recebem um *score* em função da sua posição em relação ao rótulo selecionado.

Procedimento. Uma vez extraídas as datas, é executado o módulo de avaliação que faz os cálculos de precisão, revocação e *f-measure*, medindo, assim, a qualidade da extração dos métodos CONFTRACKER e *baseline*. Além disso, avaliamos os resultados da combinação do CONFTRACKER com o *baseline*. Esta combinação foi obtida a partir dos resultados

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

[7] https://www.gnu.org/software/wget/

[8] https://github.com/zhongkaifu/CRFSharp

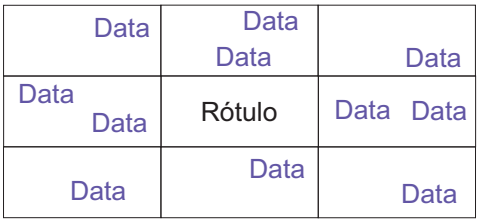

Figura 6. Posições de datas em relação a um rótulo

do CONFTRACKER, preenchendo-se as lacunas com as datas encontradas pelo *baseline*. Esta estratégia foi escolhida, pois conforme resultados apresentados e discutidos na Seção 5.2, o CONFTRACKER foi bastante superior ao *baseline*, tendo-se assim a extração do CONFTRACKER como resultado principal, sendo complementado pelo *baseline*.

5.2. Resultados e Discussão

A Tabela 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 de todas as datas de interesse, O CONFTRACKER obteve a melhor precisão (0,804) e a 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 significativa do CONFTRACKER. Isso é comprovado pelo Teste-T que resultou p-valor de $3{,}94 \times 10^{-8}$.

Comparando o CONFTRACKER com a combinação CONFTRACKER + *baseline*, observa-se uma F-measure bastante próxima. Conforme o esperado, a melhor revocação foi obtida pela combinação de métodos CONFTRACKER e *baseline* (0,694), pois desta 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 a F-measure. Assumindo que a precisão seja um índice mais importante que a revocação na resolução do problema definido neste trabalho, pois vem a ser preferível que uma data não tenha sido preenchida do que ter sido extraída erroneamente, a melhor alternativa das três abordagens aqui expostas é a utilização do método CONFTRACKER isoladamente.

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. Acreditamos que isso ocorreu o por se tratar de um intervalo de datas, distinguindo das demais *deadlines*, que são datas simples. Esta 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 esta 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 com que o CRF aprendesse a extrair esta data específica. A mesma tendência se observa com a extração do *baseline*.

Tabela 1. Resultados dos experimentos

	Abstract	Paper	Notification	Camera Ready	Conf Start	Conf End	Médias
Posicional							
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 + Posicional							
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

Limitações. Uma situação que prejudica os resultados aqui apresentados é a exibição das datas como imagens. Nestes casos não está sendo possível extrair as informações necessárias, pois no momento não fazemos uso de técnicas de *Optical Character Recognition* (OCR).

Apesar de serem previstos centenas de padrões de datas, ainda há algumas que não são contempladas, como casos em que consta o dia da semana em meio à data. Por exemplo, *04 (tue), october 2016*. Caso este 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 também a questão das datas serem atualizadas nos sites das conferências. Desta forma, faz-se necessário definir uma recorrência periódica para que as etapas (ii) e (iii) da Figura 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*, executar o processo diariamente. Assim, quando o usuário final consulta o ambiente web, a informação estará atualizada na base de dados.

6. Conclusão

Este trabalho propôs um método para a extração automática de dados de conferências. Nosso método, chamado de CONFTRACKER, emprega um processo em quatro etapas. Neste artigo, a ênfase foi na etapa de extração de datas que faz uso do *fremework* CRF.

Para realização de experimentos e avaliações, algumas conferências foram selecionadas randomicamente a partir da Tabela Qualis, sobre as quais foi gerado o *gold standard*. Para fins de comparação, implementamos um *baseline* baseado na distância entre os rótulos e as datas. Avaliamos também a combinação da técnica proposta com o *baseline*. Os resultados mostraram que a técnica proposta é significativamente melhor do que o *baseline*.

Como exposto, ambas as técnicas de extração, utilizadas para relacionar rótulos e datas dependem imensamente da capacidade de localizar as informações em linguagem

natural e não estruturada (os rótulos) utilizada nas páginas Web. Trabalhos futuros podem ser realizados no sentido de 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 são salvos em uma base de dados e ao executar a função geradora da *feature*, verifica-se a existência do *token* em questão na base de dados. Seria possível trabalhar a aplicação de aprendizagem de máquina para encontrar os *tokens* relacionados a cada data de interesse para que o valor gerado pela *feature* seja mais confiável.

Agradecimentos: Este trabalho foi parcialmente financiado pelo CNPq.

Referências

Fábio L Correia, Rui FS Amaro, Luís Sarmento, and Rosaldo JF Rossetti. Allcall: An automated call for paper information extractor. In *Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on*, pages 1–4, 2010.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

Lei Fu, Yingju Xia, Yao Meng, and Hao Yu. Conditional random fields model for web content extraction. In *Computing in the Global Information Technology (ICCGI)*, pages 30–34, 2010.

Tomas Gogar, Ondrej Hubacek, and Jan Sedivy. *Deep Neural Networks for Web Page Information Extraction*, pages 154–163. 2016.

Yunfei Gong and Qiang Liu. Automatic web page segmentation and information extraction using conditional random fields. In *Computer Supported Cooperative Work in Design (CSCWD)*, pages 334–340, 2012.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

Xinyu Li, Roya Rastan, John Shepherd, and Hye Young Paik. Automatic affiliation extraction from calls-for-papers. In *Proceedings of the Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 97–102, 2013. ISBN 978-1-4503-2411-3.

Jochen Mattes. Automated meta-data extraction for confsearch. Technical report, 2011.

Hoa Nguyen, Thanh Nguyen, and Juliana Freire. Learning to extract form labels. *Proceedings of the VLDB Endowment*, 1(1):684–694, 2008.

David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. Table extraction using conditional random fields. In *Proceedings of the annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242, 2003.

Elaine Pereira de Souza and Maria Carlota de Souza Paula. Qualis: a base de qualificação dos periódicos científicos utilizada na avaliação capes. *InfoCAPES Boletim Informativo*, 10(2), 2002.

Henry S Vieira, Altigran S da Silva, Marco Cristo, and Edleno S de Moura. A self-training crf method for recognizing product model mentions in web forums. In *European Conference on Information Retrieval*, pages 257–264, 2015.

Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma. 2d conditional random fields for web information extraction. In *Proceedings of the International Conference on Machine Learning*, pages 1044–1051, 2005.