UNIVERSIDADE FEEVALE

JONATHAN EGÍDIO SZABLEVSKI DE MOURA

EXTRAÇÃO AUTOMÁTICA DE DADOS ESTRUTURADOS DE VAGAS DE EMPREGO EM PÁGINAS WEB

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo
2018

JONATHAN EGÍDIO SZABLEVSKI DE MOURA

EXTRAÇÃO AUTOMÁTICA DE DADOS ESTRUTURADOS DE VAGAS DE EMPREGO EM PÁGINAS WEB

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

Orientador: Rodrigo Rafael Villarreal Goulart

Novo Hamburgo
2018

RESUMO

Este trabalho trata da extração automática de dados estruturados a partir de páginas HTML. Programas que extraem dados estruturados a partir de dados semi-estruturados em páginas na web são chamados de *wrappers*. Técnicas de extração automática permitem que aplicações extraiam dados sem a necessidade de intervenção humana no processo de criação dos *wrappers*. Esta pesquisa explora um dos problemas encontrados dentro do contexto de atuação da startup Jober. A Jober utiliza uma abordagem manual para o desenvolvimento de *wrappers* para coleta de dados relacionados a vagas de emprego em diferentes websites, o que limita a sua capacidade de escalar sua operação. Sendo assim, este trabalho tem como objetivo o desenvolvimento do protótipo de um *software* capaz de extrair dados estruturados a partir de páginas web contendo vagas de emprego, de modo automatizado e não-supervisionado. Para atingir este objetivo, serão identificadas as técnicas de extração automática mais relevantes para o cenário analisado, e que servem como base para o desenvolvimento do protótipo. A avaliação do protótipo será realizada a partir da comparação entre os dados extraídos com o protótipo e os dados extraídos com o *wrapper* manual já utilizado pela startup.

Palavras-chave: Processamento da Linguagem Natural, mineração de textos, wrapper.

SUMÁRIO

MOTIVAÇÃO

Atualmente, a web é a grande fonte de conhecimento da população em geral. Considerada como uma novidade promissora décadas atrás, sua relevância no dia a dia das pessoas cresceu de modo significativo, e hoje funciona como plataforma para uma grande esfera de websites e aplicações. Dentro deste contexto, é possível perceber que o acesso à imensa quantidade de dados presentes em páginas HTML possibilita a criação de aplicações ainda mais complexas e relevantes, ou, como explica Liu (2011, pág. 363, tradução nossa): "Com mais e mais empresas e organizações disseminando informações na *web*, a habilidade de extrair esses dados das páginas *web* está se tornando cada vez mais importante".

Um dos grandes desafios para a utilização dos dados presentes na web reside no fato de que o formato HTML foi concebido originalmente apenas como um meio de transmissão de dados, fazendo com que a sua natureza desestruturada dificulte consultas mais elaboradas (CRESCENZI et al., 2001). Embora existam propostas já estabelecidas com relação à utilização de dados estruturados na web, como as tecnologias de web semântica, RDF e OWL (BIZER et al., 2005), até o presente momento não há uma estimativa de curto prazo para que essas práticas sejam adotadas de maneira massiva pela comunidade de internautas.

Porém, existem muitos websites que possuem páginas "semi-estruturadas": coleções de páginas dinâmicas, que são geradas a partir de um mesmo *template* com base em dados relacionais (normalmente fornecidos por um sistema de banco de dados). Com o objetivo de explorar essa característica, diversos pesquisadores estudaram técnicas para o desenvolvimento de *wrappers*. Basicamente, um *wrapper* é um programa que realiza a extração de dados estruturados a partir de estruturas similares no HTML exibido por websites dinâmicos.

De acordo com Liu (2011), existem três abordagens principais para a construção de *wrappers*: uma abordagem manual, a indução de *wrappers* e a extração automática. Na **abordagem manual**, é necessário que o programador analise o código-fonte da página em busca dos padrões, construindo um programa que extraia os dados pertinentes. A **indução de *wrappers*** consiste na extração semi-automática de dados, onde o programa obtém regras de extração a partir de múltiplas páginas "rotuladas" manualmente, usando essas regras para extrair dados de outras páginas similares. Finalmente, a **extração automática** busca realizar a extração não-supervisionada dos dados a partir de padrões encontrados em uma ou mais páginas. As primeiras pesquisas sobre a extração automática a receberem notoriedade foram

os algoritmos RoadRunner (CRESCENZI et al., 2001) e ExAlg (ARASU et al., 2003). Desde então, pesquisadores vêm desenvolvendo novas técnicas para melhorar a qualidade de sistemas de extração de dados.

O tema deste trabalho tem como cenário o estudo de caso de um dos problemas encontrados dentro do ambiente de trabalho da startup Jober. Jober é uma empresa que tem como missão agregar vagas de emprego de inúmeros sites na internet e exibi-las aos usuários de acordo com cada perfil. Para garantir a escalabilidade do projeto, é necessário que operação da startup seja extremamente automatizada. Ou seja, o software da empresa deve ser capaz de indexar as vagas a partir de websites da internet, armazená-las e priorizá-las de acordo com o perfil de cada usuário.

Atualmente, a empresa utiliza na extração de dados de empregos uma abordagem manual, onde são desenvolvidos *wrappers* específicos para cada website. Embora funcional, esta abordagem não atende plenamente as necessidades de escalabilidade do projeto. Há a necessidade de escrita de um novo *wrapper* para cada novo website que é adicionado à plataforma, assim como é preciso reescrever esse *wrapper* no momento em que o *template* do website passar por mudanças. Mesmo as técnicas de indução de *wrappers* não se encaixam perfeitamente às necessidades do projeto, devido à dependência da intervenção humana no momento de rotular as páginas.

Em vista das dificuldades citadas acima, o tema do trabalho está voltado à abordagem da extração automática. Mais especificamente, busca-se o desenvolvimento de um protótipo que permita a extração de dados estruturados a partir de páginas web contendo vagas de emprego. Este protótipo deve ser capaz de extrair os dados sem a supervisão de uma pessoa.

Não foram encontrados trabalhos sobre extração de dados estruturados que operem exclusivamente dentro do contexto de páginas contendo vagas de emprego. Os principais estudos sobre extração automática sempre tiveram como foco o uso de técnicas generalistas, ou seja, que independem do domínio dos dados a serem extraídos. Ainda assim, técnicas presentes em trabalhos como Jindal et al. (2010) e Kayed et al. (2010) demonstram índices de acerto promissores nos seus resultados de extração de websites de diferentes domínios. Técnicas de extração automática buscam identificar e extrair registros de dados inseridos dentro de *templates* em páginas web. Essas técnicas são viáveis pois, conforme a explicação de Liu (2011, p. 382, tradução nossa):

> **Extração automática é possível porque os registros de dados (instâncias de tuplas) em um website são geralmente codificados através de um número muito pequeno de templates fixos. É possível encontrar esses templates através da mineração de padrões repetidos em múltiplos registros de dados.**

É importante citar que os registros extraídos poderão ser aninhados. Ou seja, um registro pode conter outros registros relacionados a ele. Um exemplo são listagens de produtos, onde cada produto contém uma lista de variações, como tamanho ou cor.

Este trabalho almeja contribuir em dois aspectos. O primeiro deles é devido ao fato de que o estudo de técnicas de extração automática ainda é consideravelmente relevante. Existe um grande potencial para o uso de dados estruturados presentes implicitamente em páginas web, e há muito espaço para a melhoria dos métodos atuais: identificação de disjunções ou itens opcionais, diferenciação entre conjuntos ou tuplas no HTML, marcação de nomes de atributos, integração de dados entre diferentes domínios, ente outros (LIU, 2011). Este trabalho não busca propor soluções definitivas para estes problemas, porém, sempre que possível, serão exploradas técnicas que amenizem seu impacto no cenário estudado.

A segunda contribuição refere-se à missão da startup Jober de facilitar a busca de empregos através da indexação de vagas de emprego de múltiplos websites. Um possível aperfeiçoamento do método de extração de vagas irá, invariavelmente, auxiliar a startup a se aproximar do seu objetivo e assim aproximar candidatos de possíveis oportunidades no mercado de trabalho.

OBJETIVOS

Objetivo geral

Desenvolver um protótipo de *software* que extrai dados estruturados a partir de páginas web semiestruturadas contendo vagas de emprego.

Objetivos específicos

- Identificar as técnicas de extração automática que melhor se relacionem com o cenário;
- Desenvolver o protótipo do *software* de extração a partir das técnicas identificadas;
- Avaliar o protótipo desenvolvido e interpretar resultados da validação.

METODOLOGIA

Esta pesquisa tem o objetivo de gerar um protótipo para a solução de um problema prático, que é a extração de dados estruturados a partir de páginas web. Deste modo, é possível defini-la como uma pesquisa de natureza aplicada, pois se encaixa na definição de Prodanov et al. (2013, p. 51): “[...] objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais.”.

Um dos procedimentos utilizados é a pesquisa bibliográfica. A pesquisa tem o objetivo de levantar as técnicas mais adequadas para a aplicação no cenário do trabalho. Serão pesquisados os trabalhos mais relevantes em relação ao tema da extração automática de dados estruturados. As técnicas citadas nos trabalhos mais relevantes serão utilizadas como base durante o desenvolvimento do protótipo.

O segundo procedimento utilizado neste trabalho será a pesquisa experimental. Os experimentos buscam identificar se as técnicas de extração de dados identificadas na pesquisa bibliográfica serão capazes de gerar os mesmos resultados quando aplicadas no cenário específico dos websites que contém vagas de emprego.

A avaliação do protótipo será realizada com base nos dados que são extraídos pela versão atual do *software* da empresa Jober. Atualmente, o *software* extrai os dados relevantes através de um *wrappers* manuais, a partir de páginas de listagem e páginas de detalhe indexadas de quatro websites (Indeed, Infojobs, VAGAS.com.br, Empregos.com.br).

Ao realizar a avaliação do protótipo a partir da comparação com os *wrappers* manuais, será possível verificar suas métricas de precisão e *recall*. Conforme Liu (2011), a precisão diz respeito ao percentual de itens identificados corretamente em relação ao total de itens identificados como sendo corretos. O *recall*, por sua vez, mede o percentual entre a quantidade de itens identificados corretamente em relação à quantidade de itens que deveriam ser identificados corretamente.

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Mai	Jun
Desenvolvimento do anteprojeto	X	X		
Pesquisa e avaliação de trabalhos relacionados		X	X	X
Formulação da proposta de solução		X	X	X
Redação do Relatório Final			X	X
Testes com algumas das técnicas de extração pesquisadas			X	X
Desenvolvimento do protótipo				X

Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Desenvolvimento do protótipo	X	X		
Documentação da solução implementada		X	X	X
Geração do dataset de páginas que serão utilizadas para a a avaliação		X	X	
Avaliação do protótipo		X	X	X
Documentação dos resultados obtidos na avaliação		X	X	X
Redação do Relatório Final			X	X

BIBLIOGRAFIA

ARASU, Arvind; GARCIA-MOLINA, Hector. Extracting structured data from web pages. In: **Proceedings of the 2003 ACM SIGMOD international conference on Management of data**. ACM, 2003. p. 337-348.

BIZER, Christian et al. The impact of semantic web technologies on job recruitment processes. **Wirtschaftsinformatik 2005**, p. 1367-1381, 2005.

CRESCENZI, Valter; MECCA, Giansalvatore; MERIALDO, Paolo. Roadrunner: Towards automatic data extraction from large web sites. In: **VLDB**. 2001. p. 109-118.

JINDAL, Nitin; LIU, Bing. A generalized tree matching algorithm considering nested lists for web data extraction. In: **Proceedings of the 2010 SIAM International Conference on Data Mining**. Society for Industrial and Applied Mathematics, 2010. p. 930-941.

KAYED, Mohammed; CHANG, Chia-Hui. FiVaTech: Page-level web data extraction from template pages. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 2, p. 249-263, 2010.

LIU, Bing. **Web data mining**: exploring hyperlinks, contents, and usage data. 2th ed. London, New York: Springer, 2011.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico :** métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo, RS: Feevale, 2013. 276 p. ISBN 9788577171583 Disponível em : <http://www.feevale.br/Comum/midias/8807f05a-14d0-4d5b-b1ad-1538f3aef538/E-book%20Metodologia%20do%20Trabalho%20Cientifico.pdf>. Acesso em : 26 mar. 2018.