
Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

# Extraindo Estrutura de Entidades na Web com Pouca Supervisão

Lucas Almeida Pereira de Lima

Proposta de Trabalho de Graduação

Orientador: Prof. Luciano de Andrade Barbosa

Recife
Setembro de 2016

# Resumo

A World Wide Web contém uma quantidade vasta de dados de variados domínios, e.g., imóveis, carros e produtos. Uma parte desses dados exibe uma estrutura definida representando entidades com seus atributos e respectivos valores. A obtenção de dados de entidades estruturadas pode ser útil, por exemplo, para melhoria dos resultados de busca de engenhos de busca e tomada de decisão apoiada em dados. Essa estrutura encontra-se normalmente de forma implícita, acessível apenas através de páginas HTML. O objetivo desse trabalho é desenvolver um extrator que com pouca supervisão seja capaz de extrair dados estruturados de entidades em páginas web (apartamentos, televisores etc) de diversos sites em um certo domínio.

**Palavras-chave:** recuperação de informação, extração de dados, extração de entidades, dados estruturados, coleta de dados, web

# Abstract

The World Wide Web contains a vast amount of data from a diverse set of domains, e.g., real estate, cars and products. A significant amount of this data exhibits a defined structure representing entities and their attributes and respective values. Data of web structured entities can be used, for instance, to improve search engine results, and for data-driven analytics. This structure is usually implicit, accessible only through HTML pages. The main goal of this work is to develop a semi-supervised extractor, capable of extracting structured data from entities in web pages (apartments, TVs etc).

**Keywords:** information retrieval, data extraction, entity extraction, structure data, data collection, web

# Sumário

CAPÍTULO 1

# Contexto

A World Wide Web é uma grande fonte de dados estruturados [1]. O processamento de dados estruturados de páginas web possibilita diversos benefícios, como visualização mais adequada da informação, buscas mais precisas e tomada de decisões apoiada em dados. Para processar esses dados, é preciso primeiramente extraí-los. A dificuldade da tarefa de extração de dados é dependente da forma como os dados se apresentam. Quando há uma estrutura bem definida, como em sistemas de bancos de dados, a extração pode ser feita de forma simples e precisa. Por outro lado, extrair informações de texto livre é difícil visto que não há estrutura bem definida. Como grande parte da informação da Web se apresenta de forma semi-estruturada, é possível obter uma certa estrutura através da árvore HTML e de objetos como listas e tabelas. Por outro lado, HTML é uma linguagem para objetos a serem exibidos na tela, e não para o consumo de programas de extração de dados [2].

Dessa forma, há um grande interessante em, dado um conjunto de páginas web, extrair a informação semi-estruturada relevante e transformá-la em uma "base de dados" estruturada [3]. Muitas vezes, a informação exibida em sites é gerada automaticamente através de programas que transformam a informação disponível em um banco de dados em uma página HTML. Com isso, para um mesmo site, é comum que haja um template, i.e, uma certa estrutura regular que define a forma como a informação é exibida na página [4].

Uma abordagem simples para extração de dados estruturados consiste em analisar páginas de um dado site para identificar características específicas desse site (como caminho na árvore DOM e atributos do HTML) para extração. Essa estratégia é capaz de alcançar uma boa acurácia, porém, precisa de um grande esforço manual para repetir o processo em muitos sites, além do custo de manter os extratores atualizados quando ocorrem mudanças na estrutura. Dessa forma, deseja-se neste trabalho criar para um determinado domínio um extrator único para extrair dados estruturados de vários sites [5][6][7][8][9][10]. O grande desafio de desempenhar essa tarefa decorre da grande variedade na forma de exibição de conteúdo encontrada entre sites, até mesmo dentro de um mesmo domínio.

CAPÍTULO 2

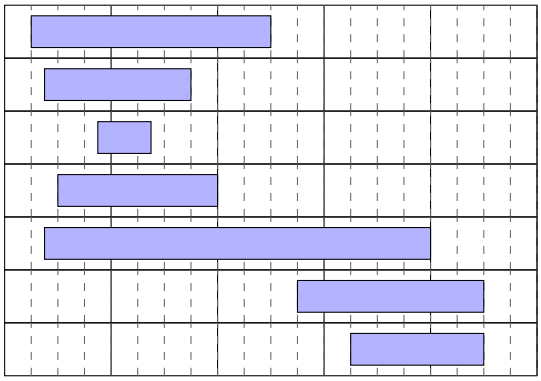
# Objetivo

O objetivo principal desse trabalho é a criação de um extrator semi-supervisionado para dados estruturados em páginas HTML em determinado domínio, que seja efetivo para uma quantidade extensa de sites. Como estudo de caso será feita a extração de dados estruturados de entidades do domínio de imóveis.

CAPÍTULO 3

# Cronograma

| Atividades | Ago | Set | Out | Nov | Dez |
|---|---|---|---|---|---|
| Leitura Bibliográfica | | | | | |
| Coleta e análise de páginas | | | | | |
| Elaboração da proposta | | | | | |
| Implementação de extratores específicos | | | | | |
| Implementação e avaliação do extrator genérico | | | | | |
| Criação da demonstração | | | | | |
| Escrita do relatório final | | | | | |



CAPÍTULO 4

# Possíveis Avaliadores

São possíveis avaliadores do trabalho a ser produzido conforme especificado nesta proposta:

- Bernadette Farias Lóscio
- Ana Carolina Salgado

CAPÍTULO 5

# Assinaturas

---

Lucas Almeida Pereira de Lima
Aluno

---

Luciano de Andrade Barbosa
Orientador

# Referências Bibliográficas


[1] N. Dalvi, A. Machanavajjhala, and B. Pang, "An analysis of structured data on the web," *Proc. VLDB Endow.*, vol. 5, pp. 680–691, Mar. 2012.

[2] T. Weninger and J. Han, "Information network analysis and extraction on the world wide web." Proceedings of the 2013 International Conference on World Wide Web (WWW 2013), May 2013.

[3] M. J. Cafarella, A. Halevy, and J. Madhavan, "Structured data on the web," *Commun. ACM*, vol. 54, pp. 72–79, Feb. 2011.

[4] H. G.-M. Arvind Arasu, "Extracting structured data from web pages," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, Inc., June 2003.

[5] L. Barbosa and G. Ferreira, *Extracting Records and Posts from Forum Pages with Limited Supervision*, pp. 233–240. Cham: Springer International Publishing, 2015.

[6] B. Liu, R. Grossman, and Y. Zhai, "Mining data records in web pages," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, (New York, NY, USA), pp. 601–606, ACM, 2003.

[7] T. Weninger, T. J. Johnston, and J. Han, "The parallel path framework for entity discovery on the web," *ACM Trans. Web*, vol. 7, pp. 16:1–16:29, Sept. 2013.

[8] T. Furche, G. Gottlob, G. Grasso, X. Guo, G. Orsi, C. Schallhart, and C. Wang, "Diadem: Thousands of websites to a single database," *Proc. VLDB Endow.*, vol. 7, pp. 1845–1856, Oct. 2014.

[9] D. Qiu, L. Barbosa, X. L. Dong, Y. Shen, and D. Srivastava, "Dexter: Large-scale discovery and extraction of product specifications on the web," *Proc. VLDB Endow.*, vol. 8, pp. 2194–2205, Sept. 2015.

[10] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "Webtables: Exploring the power of tables on the web," *Proc. VLDB Endow.*, vol. 1, pp. 538–549, Aug. 2008.