Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Coleta de Sites de Entidades em Larga Escala

Duhan Caraciolo Maia Souza

Trabalho de Graduação

Orientador: Prof. Luciano de Andrade Barbosa

Recife
Setembro de 2016

Resumo

A *World Wide Web* é um repositório que contém inúmeras páginas com diversos tipos de informação. Algumas dessas páginas contêm informação sobre instâncias de entidades estruturadas com seus atributos e valores associados, por exemplo, uma página que contém informação sobre um computador informando seus atributos como seu sistema operacional e seu processador. O objetivo principal deste trabalho é localizar em larga escala sites que contenham páginas de entidades estruturadas em um determinado domínio. Uma vez encontrados esses sites, uma coleta posterior dentro desses sites pode ser feita para a criação de uma base de dados no domínio definido. Para a localização desses sites, será desenvolvido um *focused crawler* que navega pela Web, explorando a vizinhança do grafo de sites já conhecidos. Para avaliar a solução proposta, foi implementado um estudo de caso no domínio de imóveis.

Palavras-chave: *focused crawler*, classificação de página, coleta de site, *backlinks*

Abstract

The World Wide Web is a repository with innumerous pages of all kind of information. Some of these pages contain information about instances of structured entities along with its attributes and associated values, for example, a webpage that contains information about a laptop showing its attributes as its operational system and processor. The primary goal of this work is to locate in large scale sites that contain webpages of structured entities of a determined domain. Once located, a follow-up gathering inside these sites can be made for the creation of a database in the specified domain. To locate these sites, a focused crawler that crawls through the web will be developed, exploring the graph neighborhood of already known sites. To evaluate the proposed solution, a case study in the real estate domain was implemented.

Keywords: focused crawler, page classification, site gathering, backlinks

Sumário

CAPÍTULO 1

Contexto

A *World Wide Web* é um repositório com bilhões de páginas[1] em diversos tópicos e em constante modificação. Muitas dessas páginas contêm conteúdo de entidades estruturadas [CHM11]. Definimos uma página da *Web* como página de entidade estruturada caso ela possua informações sobre uma determinada entidade contendo atributos e valores associados a esses atributos, por exemplo, uma página que contém características de um celular como seu preço, modelo, memória, etc. O conteúdo de entidades estruturadas na Web tem sido usado por diversas aplicações. Por exemplo, engenhos de busca têm usado esses dados para adicionar informação estruturada de uma dada consulta às suas páginas de respostas; empresas como Factual[2] têm comercializado dados estruturados coletados de páginas Web, etc.

Chamamos um site da *Web* de site de entidades caso o principal propósito dele é ter páginas de entidades, por exemplo, o site Peixe Urbano[3] possui diversas páginas de entidades para shows, restaurantes, hotéis, entre outros.

O principal objetivo deste trabalho é coletar sites de entidades na Web. Devido ao grande volume de dados na *Web*, encontrar um grande número de sites manualmente em um determinado domínio é inviável. Para isso, este trabalho propõe um coletor focado que *automaticamente* e em *larga escala* encontra sites de entidades em um determinado domínio.

Um dos desafios de se desempenhar essa tarefa é que apenas uma pequena porção da *Web* possui conteúdo relevante para um determinado domínio. Como consequência, uma estratégia de coleta que não foque no domínio pode levar o *crawler* a visitar conteúdo não relevante e consequentemente acarretar em uso inadequado de processamento e banda. *Focused crawlers* [CvdBD99][JSYL13] surgiram para resolver esse problema. Eles são *crawlers* que implementam estratégias de coleta especializadas em encontrar conteúdo no domínio de interesse.

Para este trabalho, iremos implementar um *focused crawler* que navega na Web à procura de sites de entidades estruturadas em um determinado domínio. Para isso, ele precisa desempenhar duas tarefas básicas: (1) detectar de forma efetiva e eficiente se um site visitado pelo crawler possui de fato entidades no domínio desejado; e (2) localizar sites de entidades no domínio visitando o grafo da Web.

[1] worldwidewebsize.com
[2] https://www.factual.com/
[3] https://www.peixeurbano.com.br/

CAPÍTULO 2

Objetivo

O objetivo principal deste trabalho é desenvolver uma solução para encontrar sites que contêm páginas de entidade *em larga escala*. Como um estudo de caso, a solução irá coletar URLs de sites que contêm páginas de entidade de imóveis, como por exemplo, zapimoveis[1] e olx[2].

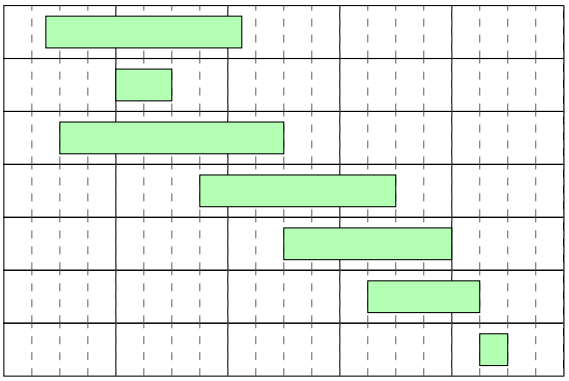
[1] http://www.zapimoveis.com.br
[2] http://www.olx.com.br/imoveis

CAPÍTULO 3

Cronograma

As atividades serão desenvolvidas conforme o cronograma abaixo, baseado no calendário acadêmico[1].

Atividades	Ago	Set	Out	Nov	Dez
Revisão Bibliográfica					
Elaboração da proposta					
Implementação do detector					
Implementação do localizador					
Escrita do relatório final					
Preparação da defesa e demo					
Defesa					



[1] https://goo.gl/zRnVc3

CAPÍTULO 4

Possíveis Avaliadores

Um possível avaliador para este trabalho é o professor Ricardo Bastos Cavalcante Prudêncio (CIn / UFPE).

CAPÍTULO 5

Assinaturas

Duhan Caraciolo Maia Souza
Aluno

Luciano de Andrade Barbosa
Orientador

Referências Bibliográficas

[CHM11] M. J. Cafarella, A. Halevy, and J. Madhavan. Structured data on the web. *Communications of the ACM*, 54(2):72–79, 2011.

[Cho01] J. Cho. *Crawling the Web: Discovery and Maintenance of a Large-Scale Web Data.* PhD thesis, Stanford University, 2001.

[CvdBD99] S. Chakrabarti, M. van den Berg, and B Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[JSYL13] J. Jiang, X. Song, N. Yu, and C. Lin. Focus: Learning to crawl web forums. *IEEE Trans. Knowl. Data Eng.*, 25(6):1293–1306, 2013.