Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Um Coletor Inteligente para Entidades Estruturadas em Sites

Bertha Maria Correia Andaluz

Proposta de Trabalho de Graduação

Orientador: Prof. Luciano de Andrade Barbosa

Recife
Setembro de 2016

Resumo

A World Wide Web é uma grande fonte de entidades estruturadas. Entidades podem ser definidas como objetos do mundo real com seus atributos e valores associados. Exemplos de entidades estruturadas na Web são especificação de produtos e infoboxes no Wikipédia. O valor prático de repositórios compreensivos de entidades estruturadas de um domínio específico é facilmente observável, por exemplo, na melhoria de resultados para consultas de entidades - o tipo mais frequente de buscas após consultas navegacionais - e decisões baseadas em dados. Antes de poderem-se utilizar esses repositórios, é necessário que essas entidades estruturadas sejam coletadas. O objetivo principal deste trabalho é implementar uma solução eficiente para o sub-problema de identificação e coleta em um mesmo website.

Palavras-chave: recuperação de informação, localização de entidades, identificação de entidades, dados estruturados, coleta de dados, web

Abstract

The World Wide Web is a great resource for structured entities. Entities can be defined as real world objects defined by their attributes and associated values. The practical value of comprehensive repositories on domain specific structured entities can easily be seen, for instance, on improving results on entity queries - the most frequent type of query search after navigational queries - and data-driven decisions. Before making use of those repositories, one must collect such structured entities. The main goal of this work is to implement an efficient solution to the sub-task of identifying and locating pages containing relevant structured entities within a website.

Keywords: information retrieval, entity location, entity identification, structure data, data collection, web

Sumário

CAPÍTULO 1

Contexto

Vem-se observando um aumento de interesse na utilização de dados estruturados na Web para diversas aplicações como integração de dados, análises estatísticas e buscas por entidades [1], que são o tipo mais frequente de buscas após consultas navegacionais [2, 3]. Neste trabalho, voltamos nossa atenção a uma parcela específica dos dados Web estruturados que chamaremos de entidades estruturadas - dados que descrevem objetos do mundo real, como produtos, lugares e pessoas, através de seus atributos e valores associados.

A tarefa de coletar e organizar em um repositório todas essas entidades estruturadas presentes na Web, independentemente do domínio a que elas pertençam, demanda uma quantidade altíssima de recursos computacionais: a obtenção de um repositório completo requeriria a análise de ao menos 4.7 bilhões [4, 5] de páginas Web. Além deste fator, a recuperação de entidades alheia ao domínio das mesmas é um empecilho para a adição de semântica [6] aos dados coletados e, consequentemente, para a extração de informação dos mesmos. Levando em conta essas considerações, optamos por restringir nosso escopo a entidades pertencentes a um domínio específico.

É fácil encontrar motivações práticas para repositórios de entidades de um domínio específico: ter informações como horário de funcionamento, meios de contato, localização e avaliações de todos os restaurantes em uma dada região; informações sobre todos os livros, suas avaliações, gêneros e autores; informações sobre artistas musicais, suas discografias e influências entre inúmeros outros possíveis exemplos.

O objetivo deste trabalho é localizar páginas de entidades estruturadas em um dado site. Esse problema possui uma escala menor em relação à de identificação de sites pertencentes ao domínio já que não é mais necessário navegar toda a Web. Todavia é de extrema importância que seja fornecida uma solução eficiente porque, para a obtenção de um repositório de entidades completo para o domínio em questão, é necessário visitar não apenas grandes agregadores de entidades, mas também sites periféricos [6]. O custo, portanto, de se visitar esse grande número de sites é bastante alto e para poder dimnuí-lo, precisam-se evitar regiões improdutivas desses sites.

CAPÍTULO 2

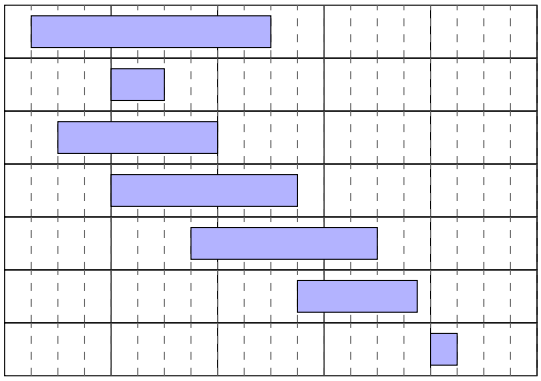
Objetivo

O objetivo deste trabalho é desenvolver um coletor intra-site, isto é, um mecanismo que localize e detecte páginas de entidades estruturadas dentro de sites de forma eficiente. O escopo deste trabalho está restrito a um domínio específico, porém pretende-se desenvolver um sistema genérico o suficiente de modo que este sistema possa ser facilmente adaptável a domínios distintos assim como servir de infraestrutura para trabalhos futuros na mesma área.

CAPÍTULO 3

Cronograma

Atividades	Ago	Set	Out	Nov	Dez
Revisão Bibliográfica					
Elaboração da proposta					
Coleta e rótulo de dados					
Desenvolvimento do classificador					
Desenvolvimento do coletor					
Relatório final					
Apresentação					



CAPÍTULO 4

Possíveis Avaliadores

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

- Nivan Roberto Ferreira Júnior
- Ricardo Bastos Cavalcante Prudêncio

CAPÍTULO 5

Assinaturas

Bertha Maria Correia Andaluz
Aluna

Luciano de Andrade Barbosa
Orientador

Referências Bibliográficas

[1] 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.

[2] F. Alahmari and L. Magee, "Linked data and entity search: A brief history and some ways ahead," in *3rd Australasian Web Conference (AWC 2015)* (J. G. Davis and A. Bozzon, eds.), vol. 166 of *CRPIT*, (Sydney, Australia), pp. 29–38, ACS, 2015.

[3] J. Pound, P. Mika, and H. Zaragoza, "Ad-hoc object retrieval in the web of data," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 771–780, ACM, 2010.

[4] M. de Kunder, "Worldwidewebsize.com | the size of the world wide web (the internet)," 2016. [Online; accessed 13-September-2016].

[5] M. de Kunder, "Geschatte Grootte van het Geïndexeerde World Wide Web," Master's thesis, Tilburg University, the Netherlands, 2006.

[6] 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.