Eduardo Teixeira Cardoso

Efficient methods for information extraction in news webpages

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Postgraduate program in Informatics

Rio de Janeiro
August 2011

Pontifícia Universidade Católica
do Rio de Janeiro

Eduardo Teixeira Cardoso

Efficient methods for information extraction in news webpages

Dissertação de Mestrado

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática, PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática

Advisor: Prof. Eduardo Sany Laber

Rio de Janeiro
August 2011

Pontifícia Universidade Católica
do Rio de Janeiro

Eduardo Teixeira Cardoso

Efficient methods for information extraction in news webpages

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática, PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the following commission:

Prof. Eduardo Sany Laber
Advisor
Departamento de Informática — PUC–Rio

Prof. Raúl Pierre Rentería
Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova
Departamento de Informática — PUC–Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro — August 24, 2011

Eduardo Teixeira Cardoso

Eduardo graduated from Pontifícia Universidade Católica do Rio de Janeiro in Information Systems. He worked in R&D projects with Microsoft Enterprise Search team in Rio de Janeiro and did an internship at Google in Belo Horizonte. His research focuses on efficient algorithms for solving problems in the web.

Bibliographic data

Cardoso, Eduardo Teixeira

Efficient methods for information extraction in news webpages / Eduardo Teixeira Cardoso ; advisor: Eduardo Sany Laber. — 2011.
58 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)-Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2011.
Inclui bibliografia

1. Informática – Teses. 2. Segmentação de notícias. 3. Renderização de páginas web. 4. Aprendizado de máquina. I. Laber, Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I'd like to thank all my friends and family for the support during these years. Even when I wasn't the most pleasant person to be around, they were there for me. Thank you very much, I couldn't have done it without you.

Abstract

Cardoso, Eduardo Teixeira; Laber, Eduardo Sany. **Efficient methods for information extraction in news webpages**. Rio de Janeiro, 2011. 58p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

We tackle the task of news webpage segmentation, specifically identifying the news title, publication date and story body. While there are very good results in the literature, most of them rely on webpage rendering, which is a very time-consuming step. We focus on scenarios with a high volume of documents, where a short execution time is a must. The chosen approach extends our previous work in the area, combining structural properties with hints of visual presentation styles, computed with a faster method than regular rendering, and machine learning algorithms. In our experiments, we took special attention to some aspects that are often overlooked in the literature, such as processing time and the generalization of the extraction results for unseen domains. Our approach has shown to be about an order of magnitude faster than an equivalent full rendering alternative while retaining a good quality of extraction.

Keywords

News segmentation. Webpage rendering. Machine learning.

Resumo

Cardoso, Eduardo Teixeira; Laber, Eduardo Sany. **Métodos eficientes para extração de informação em páginas de notícias**. Rio de Janeiro, 2011. 58p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Nós abordamos a tarefa de segmentação de páginas de notícias; mais especificamente identificação do título, data de publicação e corpo da notícia. Embora existam resultados muito bons na literatura, a maioria deles depende da renderização da página, que é uma tarefa muito demorada. Nós focamos em cenários com um alto volume de documentos, onde desempenho de tempo é uma necessidade. A abordagem escolhida estende nosso trabalho prévio na área, combinando propriedades estruturais com traços de atributos visuais, calculados através de um método mais rápido do que a renderização tradicional, e algoritmos de aprendizado de máquina. Em nossos experimentos, nos atentamos para alguns fatos não comumente abordados na literatura, como tempo de processamento e a generalização dos nossos resultados para domínios desconhecidos. Nossa abordagem se mostrou aproximadamente uma ordem de magnitude mais rápida do que alternativas equivalentes que se apoiam na renderização completa da página e manteve uma boa qualidade de extração.

Palavras–chave

Segmentação de notícias. Renderização de páginas web. Aprendizado de máquina.

Contents

List of Figures

List of Tables

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke, *Profiles of the Future.*