

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



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



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

All rights reserved.

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

1	Introduction	11
1.1	Related work	12
1.2	Our contribution	14
1.3	Organization	15
2	Important concepts	16
2.1	HyperText Markup Language (HTML)	16
2.2	Document Object Model (DOM)	17
2.3	Cascading Style Sheets (CSS)	18
2.4	Machine learning methods used	19
2.5	Metrics	21
3	Our approach	24
3.1	Simplified CSS parser	24
3.2	Attributes	25
3.3	Training process	27
3.4	Classification process	27
4	Evaluation of selected works in the literature	31
4.1	Luo, et al. (2009)	31
4.2	Wang, et al. (2009)	33
5	Experiments	36
5.1	Corpora	36
5.2	Evaluating works in the literature	41
5.3	Our approach	45
6	Conclusion	51
	Bibliography	52
A	List of sites in the RCD4 corpus	57

List of Figures

1.1	Illustration of the regions of interest in a news webpage. The rectangles identify the date, title and body, respectively numbered 1, 2 and 3.	12
2.1	Example of an HTML document.	16
2.2	Example of a DOM tree representation of HTML documents	17
2.3	Example of a style sheet document.	18
2.4	Example of a document's presentation before and after applying some CSS rules.	19
2.5	Example of a decision tree to decide if the weather is good to play outside.	20
2.6	Illustration of a hyperplane that linearly separates points from two distinct classes.	21
2.7	Illustration of how the kernel trick can affect the data points.	21
3.1	Outline of our execution pipeline	24
3.2	Example of a DOM subtree that will receive preprocessing.	27
4.1	Example of using DOM text nodes versus text segments.	32
4.2	Examples of non-obvious cases for the postprocessing step of [22].	33
4.3	Illustration of the dimensions considered for the vertical distance and horizontal overlap features for title nodes.	34
4.4	Illustration of potential issues with their body node classification; notice how the node that contains the news body also contains unimportant content.	35
5.1	Example of a potential weakness of annotating element nodes.	40
5.2	Example of a shared weakness of element node and text node annotations.	40
5.3	Screenshot of our tool for visual annotation of webpages.	41

List of Tables

5.1	Results of our implementation of Luo, et al. (2009) for full and partial rendering approaches.	42
5.2	Results for body detection.	44
5.3	Results for title detection.	45
5.4	Basis of comparison for our approach; results from our implementation of Wang et al. (2009) [40, 41] on the Cardoso3000 corpus, measured with a content-based bag of words metric.	45
5.5	Title extraction results on the RCD4 corpus for strictly structural attributes	46
5.6	Title extraction results on the RCD4 corpus for both structural and visual attributes	46
5.7	Title extraction execution times on the RCD4 corpus, relative to strictly structural approach	47
5.8	Date extraction results on the RCD4 corpus for the conditional approach with two models, after post-processing	47
5.9	Date extraction results on the RCD4 corpus for the title-dependent model, after post-processing	47
5.10	Body extraction results on the RCD4 corpus	48
5.11	Results validation for the task of title detection.	48
5.12	Results validation for the task of body detection.	49
5.13	Results validation for the task of date detection.	49
5.14	Results validation for the task of date detection.	50
5.15	Results on the NEWS600 corpus, ordered by F1. The baseline and our results are measured using bag of words.	50

Any sufficiently advanced technology is indistinguishable from magic.

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