Pontifícia Universidade Católica
do Rio de Janeiro

Evelin Carvalho Freire de Amorim

NCE: Um algoritmo para extração de conteúdo de páginas de notícias

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio

Orientador: Prof. Eduardo Sany Laber

Rio de Janeiro
Março de 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Evelin Carvalho Freire de Amorim

NCE: Um algoritmo para extração de conteúdo de páginas de notícias

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

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

Prof. Raúl Pierre Rentería
Departamento de Informática - PUC-Rio

Prof. Ruy Luiz Milidiú
Departamento de Informática - PUC-Rio

Prof. José Eugênio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 31 de Março de 2009

Evelin Carvalho Freire de Amorim

Ficha Catalográfica

Amorim, Evelin Carvalho Freire

NCE: um algoritmo para extração de conteúdo de páginas de notícias / Evelin Carvalho Freire de Amorim; orientador: Eduardo Sany Laber. : 2009.

83 f: il. ; 30 cm

1. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2009.

Inclui bibliografia.

1. Informática – Tese. 2. Extração. 3. Web. 4. Algoritmos. 5. Árvores. 6. Máquinas de Busca. I. Laber, Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Agradeço primeiramente a Deus, por verdadeiramente reconhecer que sem Sua ajuda e cuidado as vitórias não aconteceriam.

Ao meu orientador Eduardo Laber pelo incentivo para a realização deste trabalho

À Capes, a Faperj e a PUC–Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

À minha mãe Rosangela pelo apoio incondicional, à família pelo incentivo e ao Leonardo Contadini pelo compreensão.

Aos amigos do LEARN, especialmente Críston, Eduardo Cardoso e Iam Jabour, que me auxiliaram com explicações, orientações técnicas e com a amizade.

Ao pessoal do departamento de Informática para a ajuda de todos os dias.

Resumo

Amorim, Evelin Carvalho Freire; Laber, Eduardo Sany. **NCE: Um algoritmo para extração de conteúdo de páginas de notícias**. Rio de Janeiro, 2009. 83p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A extração de entidades de páginas *web* é comumente utilizada para melhorar a qualidade de muitas tarefas realizadas por máquinas de busca como detecção de páginas duplicadas e *ranking*. Essa tarefa se torna ainda mais relevante devido ao crescente volume de informação da internet com as quais as máquinas de busca precisam lidar. Existem diversos algoritmos para detecção de conteúdo na literatura, alguns orientados a *sites* e outros que utilizam uma abordagem mais local e são chamados de algoritmos orientados a páginas. Os algoritmos orientados a *sites* utilizam várias páginas de um mesmo *site* para criar um modelo que detecta o conteúdo relevante da página. Os algoritmos orientados a páginas detectam conteúdo avaliando as características de cada página, sem comparar com outras páginas. Neste trabalho apresentamos um algoritmo, chamado `NCE` (*News Content Extractor*), orientado a página e que se propõe a realizar extração de entidades em páginas de notícias. Ele utiliza atributos de uma árvore DOM para localizar determinadas entidades de uma página de notícia, mais especificamente, o título e o corpo da notícia. Algumas métricas são apresentadas e utilizadas para aferir a qualidade do `NCE`. Quando comparado com outro método baseado em página e que utiliza atributos visuais, o `NCE` se mostrou superior tanto em relação à qualidade de extração quanto no que diz respeito ao tempo de execução.

Palavras–chave

Extração; Web; Algoritmos; Árvores; Máquinas de Busca;

Abstract

Amorim, Evelin Carvalho Freire; Laber, Eduardo Sany(Advisor). **NCE: An algorithm for content extraction in news pages**. Rio de Janeiro, 2009. 83p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The entity extraction of web pages is commonly used to enhance the quality of tasks performed by search engines, like duplicate pages and ranking. The relevance of entity extraction is crucial due to the fact that search engines have to deal with fast growning volume of information on the web. There are many algorithms that detect entities in the literature, some using site level strategy and others using page level strategy. The site level strategy uses many pages from the same site to create a model that extracts templates. The page level strategy creates a model to extract templates according to features of the page. Here we present an algorithm, called `NCE` (News Content Extractor), that uses a page level strategy and its objective is to perform entity extraction on news pages. It uses features from a DOM tree to search for certain entities, namely, the news title and news body. Some measures are presented and used to evaluate how good `NCE` is. When we compare `NCE` to a page level algorithm that uses visual features, `NCE` shows better execution time and extraction quality.

Keywords

Extraction; Web; Algorithms; Trees; Search Engines;

Sumário

1 Introdução **11**
1.1 Motivação 11
1.2 Definição do Problema 12
1.3 Resultados 13
1.4 Organização da Dissertação 13

2 Pesquisa Bibliográfica **15**
2.1 Estratégias orientadas a *site* 15
2.2 Estratégias orientadas a páginas 19

3 Algoritmo NCE **25**
3.1 Busca do nó separador 25
3.2 Refinando a solução 29
3.3 Complexidade 31

4 Experimentos **34**
4.1 Métricas 34
4.2 Ambiente 36
4.3 Preparação do Corpus 36
4.4 Experimentos 37

5 Conclusão **45**

Referências Bibliográficas **47**

A Lista de Portais de Notícias **49**

B Resultados dos algoritmos NCE e V-Wrapper por página **52**

C Tempos dos algoritmos NCE e V-Wrapper por página **70**

Lista de figuras

1.1 Exemplo de template em um página de notícia 12

2.1 Exemplo de árvore DOM para ST 16
2.2 Exemplo dos atributos de blocos 23

3.1 Região que pode ser acrescentada como conteúdo relevante 27
3.2 Corpo e título de uma notícia 28
3.3 Comentários da notícia 29
3.4 Esquema da árvore DOM do documento das Figuras 3.2 e 3.3 29

4.1 Histograma de Tempo do NCE no corpus CE 38

Lista de tabelas

2.1 Atributos utilizados no algoritmo V-Wrapper 22

3.1 Valores padrão dos parâmetros do algoritmo NCE 32

4.1 Resultados para o algoritmo NCE. 37
4.2 Resultados para o algoritmo NCE sem o módulo de busca por título. 37
4.3 Resultados para o algoritmo NCE sem os módulos de remoção de comentários e busca por título. 38
4.4 Resultado do NCE por *site* do conjunto CA 39
4.5 Tempo para o algoritmo NCE no corpora 40
4.6 Lista de Testes para o algoritmo V-Wrapper 41
4.7 Validação cruzada para o algoritmo V-Wrapper 41
4.8 Resultados para o algoritmo V-Wrapper 41
4.9 Resultado V-Wrapper no corpus CA 42
4.10 Tempo para o algoritmo V-Wrapper 43

A.1 Portais utilizados na exploração - Parte 1 49
A.2 Portais utilizados na exploração - Parte 2 50
A.3 Portais utilizados nos testes 51

B.1 Resultado do V-Wrapper no Conjunto CA sem classificador de Folhas - Parte 1 52
B.2 Resultado do V-Wrapper no Conjunto CA sem classificador de Folhas - Parte 2 53
B.3 Resultado do V-Wrapper no Conjunto CA sem classificador de Folhas - Parte 3 54
B.4 Resultado do V-Wrapper no Conjunto CA sem classificador de Folhas - Parte 4 55
B.5 Resultado do V-Wrapper no Conjunto CA com classificador de Folhas - Parte 1 56
B.6 Resultado do V-Wrapper no Conjunto CA com classificador de Folhas - Parte 2 57
B.7 Resultado do V-Wrapper no Conjunto CA com classificador de Folhas - Parte 3 58
B.8 Resultado do V-Wrapper no Conjunto CA com classificador de Folhas - Parte 4 59
B.9 Resultado do V-Wrapper no Conjunto CE sem classificador de Folhas - Parte 1 60
B.10 Resultado do V-Wrapper no Conjunto CE sem classificador de Folhas - Parte 2 61
B.11 Resultado do V-Wrapper no Conjunto CE com classificador de Folhas - Parte 1 62
B.12 Resultado do V-Wrapper no Conjunto CE com classificador de Folhas - Parte 2 63
B.13 Resultado do NCE no Conjunto CA - Parte 1 64

B.14 Resultado do NCE no Conjunto CA - Parte 2 65
B.15 Resultado do NCE no Conjunto CA - Parte 3 66
B.16 Resultado do NCE no Conjunto CA - Parte 4 67
B.17 Resultado do NCE no Conjunto CE - Parte 1 68
B.18 Resultado do NCE no Conjunto CE - Parte 2 69

C.1 Tempo(s) de execução do NCE no corpus CE - Parte 1 70
C.2 Tempo(s) de execução do NCE no corpus CE - Parte 2 71
C.3 Tempo(s) de execução do NCE no corpus CA - Parte 1 72
C.4 Tempo(s) de execução do NCE no corpus CA - Parte 2 73
C.5 Tempo(s) de execução do NCE no corpus CA - Parte 3 74
C.6 Tempo(s) de execução do NCE no corpus CA - Parte 4 75
C.7 Tempo(s) de execução do NCE no corpus CA - Parte 5 76
C.8 Tempo(s) de execução do V-Wrapper no corpus CE - Parte 1 77
C.9 Tempo(s) de execução do V-Wrapper no corpus CE - Parte 2 78
C.10 Tempo(s) de Execução do V-Wrapper no corpus CA - Parte 1 79
C.11 Tempo(s) de Execução do V-Wrapper no corpus CA - Parte 2 80
C.12 Tempo(s) de Execução do V-Wrapper no corpus CA - Parte 3 81
C.13 Tempo(s) de Execução do V-Wrapper no corpus CA - Parte 4 82
C.14 Tempo(s) de Execução do V-Wrapper no corpus CA - Parte 5 83