Alexander Arturo Mera Caraballo

Publishing Annotated Multimedia Deep Web Data

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO
EM INFORMÁTICA

Rio de Janeiro
April 2012

Pontifícia Universidade Católica
do Rio de Janeiro

Alexander Arturo Mera Caraballo

Publishing Annotated Multimedia Deep Web Data

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro, April 2012.

Pontifícia Universidade Católica
do Rio de Janeiro

Alexander Arturo Mera Caraballo

Publishing Annotated Multimedia Deep Web Data

Dissertation presented to the Programa de Pós-graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática

Prof. Marco Antonio Casanova
Orientador
Departamento de Informática -- PUC-Rio

Prof. Antonio Luz Furtado
Departamento de Informática -- PUC-Rio

Prof. Luiz André Portes Paes Leme
Instituto de Computação -- UFF

Prof. José Eugenio Leal
Coordenador (a) Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, April 03, 2012.

Alexander Arturo Mera Caraballo

Graduated in Systems Engineering from Universidad de Nariño (UDENAR), Pasto – Colombia in 2009.

Bibliographic data

Mera Caraballo, Alexander Arturo

Publishing annotated multimedia Deep Web data/ Alexander Arturo Mera Caraballo; Advisor: Marco Antonio Casanova – 2012.

67f.: il. (Color) ; 30 cm

Dissertação de (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.

Inclui bibliografia

1. Informática – Teses. 2. Recuperação de conteúdo falado. 3. Recuperação de dados multimídia. 4. Objetos de aprendizagem. 5. Navegação de áudio. 6. Recuperação de informação multilíngue. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I wish to thank, first and foremost, my advisor Prof. Marco Antonio Casanova, whose encouragement, guidance and everyday kindness made possible the realization of this work.

I also would like to make a special reference to MSc. Bernardo Pereira Nunes for his friendship, valuable insights and recommendations.

I want to thank to my beloved family, for their patience and support.

I remain indebted to many colleagues and professors at PUC-Rio for providing me the means to learn and understand.

To CAPES, for the financial support, without which this work would not have been possible

Abstract

Mera Caraballo, Alexander Arturo; Casanova, Marco Antonio. **Publishing Annotated Multimedia Deep Web Data.** Rio de Janeiro, 2012. 67p. MSc Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In recent years, we witnessed a huge growth of multimedia data on the Web. New lower-cost technologies and greater bandwidth allowed the Web to evolve into a multimedia format. However, the lack of tools that can make multimedia format easily accessible on the Web led us to a non-searchable and non-indexable data of the Web, also known as Deep Web. In line with these observations, this dissertation addresses the problem of how to publish audio and video content on the Web. We present a tool and a novel approach that facilitates the indexing and retrieval of the objects with the help of traditional search engines. The tool automatically generates static Web pages that describe the content of the objects and organize this content to facilitate locating segments of the audio or video which correspond to the descriptions. The static Web pages can be translated to others languages to reach other user populations. An annotation process is also performed to embed machine-readable data into the Web pages. The dissertation also presents an in-depth experiment, publishing learning objects based on audio and video, to assess the efficacy of the technique.

Keywords

Spoken content retrieval; Multimedia retrieval; Learning objects; Audio browsing; Cross-Language Information Retrieval.

Sumário

Mera Caraballo, Alexander Arturo; Casanova, Marco Antonio. **Publicando anotações de dados multimídia advinda da Deep Web.** Rio de Janeiro, 2012. 67p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Nos últimos anos, temos assistido um enorme crescimento de dados multimídia na Web. Novas tecnologias de menor custo e maior largura de banda têm permitido que a Web evolua para um formato multimídia. No entanto, a falta de ferramentas que podem tornar o formato multimídia disponível na Web nos levou a um conjunto de dados não-pesquisável e não indexável da Web, também conhecido como Deep Web. Desta forma, esta dissertação aborda o problema de como publicar conteúdo de áudio e vídeo na Web. Apresentamos uma ferramenta e uma nova abordagem que facilita a indexação e recuperação dos objetos com a ajuda das maquinas de busca tradicionais. A ferramenta gera automaticamente páginas Web estáticas que descrevem o conteúdo dos objetos e organizar esse conteúdo para facilitar a localização de segmentos do áudio ou vídeo que correspondem às descrições. As páginas Web estáticas podem ser traduzidos para outras línguas para atingir outras populações de usuários. Um processo de anotação também é realizado para incorporar dados legíveis pelas máquinas nas páginas Web. A dissertação também apresenta um experimento completo, publicando objetos de aprendizagem baseados em áudio e vídeo para avaliar a eficácia da abordagem.

Palavras-chave

Recuperação de conteúdo falado; Recuperação de dados multimídia; Objetos de Aprendizagem; Navegação de Áudio; Recuperação de Informação Multilíngue.

Contents

List of Figures

List of Tables