1. Record Nr. UNINA9910782282803321

Titolo	Web document analysis [[electronic resource]] : challenges and opportunities / / editors, Apostolos Antonacopoulos, Jianying Hu
Pubbl/distr/stampa	Singapore ; ; River Edge, NJ, : World Scientific, 2003
ISBN	1-281-92809-7 9786611928094 981-277-537-4
Descrizione fisica	1 online resource (346 p.)
Collana	Series in machine perception and artificial intelligence ; ; v. 55
Altri autori (Persone)	AntonacopoulosApostolos HuJianying <1966->
Disciplina	005.741
Soggetti	Data mining Internet searching
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Description based upon print version of record.
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	CONTENTS; PREFACE; Part I. Content Extraction and Web Mining; CHAPTER 1 CLUSTERING OF WEB DOCUMENTS USING A GRAPH MODEL; 1. Introduction; 2. Graphs: Formal Notation; 3. The Extended k-Means Clustering Algorithm; 4. Clustering of Web Documents using the Graph Model; 5. Experimental Results; Acknowledgments; References; CHAPTER 2 APPLICATIONS OF GRAPH PROBING TO WEB DOCUMENT ANALYSIS; 1. Introduction; 2. Related Work; 3. A Formalism for Graph Probing; 4. Experimental Evaluation; 4.1. Graph Model; 4.2. Generating ""Random"" Collections of Web Pages; 4.3. Experiment #1: Full Graph Matching 4.4. Experiment #2: Subgraph Matching 5. Conclusions; 6. Acknowledgments; References; CHAPTER 3 WEB STRUCTURE ANALYSIS FOR INFORMATION MINING; 1. Introduction; 2. Object Model Architecture; 2.1. HTML Parsing Library; 2.2. Single-Slot HTML Parsing Functions; 2.3. Multi-Slot/Pattern HTML Parsing Functions; 3. User Interface; 4. News Article Extraction; 5. Link Extraction; 6. Stock Quote Extraction; 7. Conclusion; Acknowledgment; References; CHAPTER 4 NATURAL LANGUAGE PROCESSING FOR WEB DOCUMENT ANALYSIS; 1. Introduction; 2. Design Principles; 2.1. Why XML?; 2.2. User Orientation

2.3. Portability 3. Document Suite XDOC; 3.1. Preprocessing Module; 3.1.1. HTML Cleaner; 3.1.2. Structure Tagger; 3.1.3. POS Tagger; 3.2. Syntactic Module; 3.2.1. Syntactic Parser; 3.2.2. Phrase Detector; 3.3. Corpus Based Module; 3.4. Semantic Module; 3.4.1. Semantic Tagger; 3.4.2. Case Frame Analysis; 3.4.3. Semantic Interpretation of Syntactic Structure; 4. Related Work; 5. Conclusion; References; Part II. Document Analysis for Adaptive Content Delivery; CHAPTER 5 REFLOWABLE DOCUMENT IMAGES; 1. Introduction; 2. Image and Layout Analysis; 2.1. Text/Image Segmentation; 2.2. Preprocessing
2.3. Layout Analysis 3. HTML-Based Representations; 4. Reader Applications; 5. New Document Formats; 6. Summary and Conclusions; Acknowledgments; References; CHAPTER 6 EXTRACTION AND MANAGEMENT OF CONTENT FROM HTML DOCUMENTS; 1. Introduction; 2. Research Direction; 3. Current State of the Art; 3.1 Handcrafting; 3.2 Transcoding; 3.3 Adaptive Re-authoring; 4. Proposed Approach; 4.1. Web Page Segmentation; 4.2. Contextual Analysis and Segment Labeling; 4.3. Web-Page Summarization; 4.4. Post-processing; 4.5. Overall Summary of the Content Extraction and Display Process; 5. Results; 6. Discussion
6.1 Web Page Segmentation 6.2 Contextual Analysis and Segment Labeling; 6.3 Web-Page Summarization; 6.4. Display Capabilities; 6.5. Language Independence; 6.6. Current State of Research; 6.7. Supported Devices; 7. Concluding Remarks; References; CHAPTER 7 HTML PAGE ANALYSIS BASED ON VISUAL CUES; 1. Introduction; 1.1. Document Analysis for Search Engines; 1.2. Document Analysis for Adaptive Content Delivery; 2. Visual Similarity of HTML Objects; 2.1. Visual Similarity of Simple Objects; 2.2. Visual Similarity of Container Objects; 3. Pattern Detection and Construction of Structured Documents
3.1. Quantization

Sommario/riassunto

This book provides the first comprehensive look at the emerging field of web document analysis. It sets the scene in this new field by combining state-of-the-art reviews of challenges and opportunities with research papers by leading researchers. Readers will find in-depth discussions on the many diverse and interdisciplinary areas within the field, including web image processing, applications of machine learning and graph theories for content extraction and web mining, adaptive web content delivery, multimedia document modeling and human interactive proofs for web security.