

Nabil R. Adam Bharat K. Bhargava
Yelena Yesha (Eds.)

Digital Libraries

Current Issues

Digital Libraries Workshop DL '94
Newark, NJ, USA, May 19-20, 1994
Selected Papers

Technische Hochschule Darmstadt	
FACHBEREICH INFORMATIK	
B I B L I O T H E K	
Inventar-Nr.:	21218
Sachgebiete:
Standort:



Springer

Contents

1 Overview	1
by Nabil R. Adam, Bharat K. Bhargava and Yelena Yesha	
I Introduction	7
2 Some Key Issues in Database Systems in a Digital Library Setting	9
by Nabil R. Adam, Bradley S. Fordham and Yelena Yesha	
2.1 Motivation	9
2.2 Meta-data vs. Data and Self-Description	11
2.3 Support for Multiple Views	13
2.4 DL Query Languages	13
2.5 Database Utilities	14
2.6 Transaction Management	15
2.7 Files and Indexes	16
2.8 Security and Authorization	16
2.9 Database Design	17
2.10 Query Processing and Optimization	18
2.11 Database Constraints	19
3 Promising Research Direction in Digital Libraries	21
by Nabil R. Adam, Milton Halem and Shamim Naqvi .	
3.1 Motivation	21
3.2 Digital Libraries As Distributed Databases	23
3.3 Digital Libraries As Networked Information Systems	25
3.4 New use of Content	26
3.5 New Uses	28
3.5.1 Multimodal Presentations	28
3.5.2 Adaptive/Opinion indexing	28
3.5.3 Generating New Content	29
3.6 Conclusion	29

II Administration/Management	31
4 Which Way to the Future? The Control of Scholarly Publication	33
by Michael Lesk	
4.1 Introduction	33
4.2 Technical issues	35
4.3 Our experiments	37
4.4 Economic issues	45
4.5 Conclusions	48
4.6 Acknowledgments	48
5 Networked Information Systems As Digital Libraries	51
by Jacob Slonim and Lisa BaronLaurie	
5.1 Introduction and Motivation	51
5.2 Challenges	53
5.2.1 Technological Issues	53
5.2.2 Human-Centric Issues	57
5.3 Socio-Economic Issues	58
5.4 Conclusion	59
5.5 Acknowledgements	60
5.6 Trademarks	60
III Information Retrieval/Hypertext	63
6 Automatic Hypertext Conversion of Paper Document Collections	65
by Andreas Myka and Ulrich Guntzer	
6.1 Introduction	65
6.2 Preprocessing of Documents	67
6.2.1 Optical Scanning and OCR Processing	68
6.2.2 Layout Analysis	69
6.3 Link Generation	71
6.3.1 Link Description	71
6.3.2 Link Recognition	78
6.4 Hypertext Evaluation	84
6.4.1 Link Learning	84
6.4.2 Evaluation of Introduced Links	85
6.5 User Interface	85
6.5.1 HyperFacs	86
6.5.2 Conversion of hypertext bases	89
6.6 Conclusion	90

7	Administering Structured Documents in Digital Libraries	91
	by Klemens Böhm, Karl Aberer and Erich Neuhold	
7.1	Introduction	91
7.2	Review of Basic SGML Concepts	95
7.3	Key Concepts of the VODAK Modeling Language (VML)	97
7.4	A VODAK Application Framework for SGML Documents	101
7.4.1	Modeling Issues	101
7.4.2	Functionality and Experiences	104
7.5	Extending the Framework with HyTime Features	109
7.6	Embedding the Application Framework in Various Scenarios	112
7.6.1	Coupling with Information-Retrieval Systems	112
7.6.2	Combining Documents' Content with Knowledge Bases	113
7.6.3	Handling HTML Documents	113
7.6.4	Coupling with DFR-Archive	115
7.6.5	Architecture	115
7.7	Conclusions	115
8	Document Recognition for a Digital Library	119
	by Sargur N. Srihari, Stephen W. Lam and Jonathan J. Hull	
8.1	Introduction	119
8.2	Adaptive Document Layout Understanding	120
8.2.1	Skew Correction	120
8.2.2	Block Segmentation and Classification	120
8.2.3	Layout Understanding	121
8.3	Text Recognition Using Document Context	123
8.3.1	Experimental Investigation	124
8.4	Logical Linking	127
8.5	Conclusions	128
9	Using Non-Textual Cues for Electronic Document Browsing	129
	by Daniela Rus and Kristen Summers	
9.1	Introduction	129
9.1.1	Outline	131
9.1.2	Previous Work	132
9.2	The Segmentation Algorithm	133
9.2.1	The Algorithm for Generating the Logical Hierarchy	133
9.2.2	Indentation Alphabets	135
9.2.3	The Indentation Pattern Language	141
9.2.4	A Logical Hierarchy Example	143
9.2.5	Future Extensions	143
9.3	Classification Algorithms	145
9.3.1	A Classifier for Tables	145
9.3.2	A Classifier for Line Drawings	150
9.3.3	A Classifier for Itemized Lists	152
9.3.4	Future Extensions	153

9.4	Experiments	154
9.4.1	Segmentation Experiments	154
9.4.2	The Results	154
9.4.3	Application to Information Access	159
9.5	Discussion	161

IV Classification And Indexing 163

10 Corpus Linguistics for Establishing the Natural Language Content of Digital Library Documents 165

by **Robert P. Futrelle, Xiaolan Zhang and Yumiko Sekiya**

10.1	Introduction	165
10.2	Information retrieval and browsing	166
10.3	Start-up and the steady state	166
10.4	The Linguistic Database	167
10.5	Domain and genre	168
10.6	Words	168
10.7	Word classification	169
10.8	The Balanced Entropy Method	172
10.9	Word sense alignment between independent systems	174
10.10	Higher-order analysis of language	174
10.11	More complex needs — knowledge frames	175
10.12	Derivative objects	175
10.13	Diagrams — Contents and analysis	176
10.14	Multidatabases for natural language	176
10.15	Authoring tools for capturing content	176
10.16	Conclusions	177
10.17	Acknowledgments	178

11 Compression and Full-Text Indexing for Digital Libraries 181

by **Ian H. Witten, Alistair Moffat and Timothy C. Bell**

11.1	Introduction	181
11.2	The information explosion	183
11.3	Designing document databases	185
11.4	Storing the documents	189
11.5	Indexing a document collection	193
11.6	Querying a full-text information base	196
11.7	The problem of dynamic collections	199
11.8	Conclusion	200

12 The Digital Library and The Home-based User	203
by Andy Sloane	
12.1 Introduction	203
12.2 The digital library, multimedia and communications	203
12.3 Models of multimedia information provision	204
12.4 The home-based user	206
12.5 Navigation, classification, indexing and hypertext	207
12.6 Summary	208
13 Integrating Natural Language With Large Dataspace Visualization	209
by Ira Smotroff, Lynette Hirschman and Samuel Bayer	
13.1 Information Overload	209
13.2 The WAIS Information Exploration System	210
13.3 Visualization of Document Space	211
13.3.1 Architecture for the Visualization Interface	211
13.3.2 The Visualizer	212
13.3.3 Extensions to the Visualization Interface	213
13.4 Integrating Natural Language With Visualization	214
13.4.1 Information Extraction	214
13.4.2 Natural Language Interface	216
13.4.3 Visualization with Natural Language Processing	217
13.4.4 Document Summarization	218
13.5 Future Directions	219
13.6 Acknowledgements	220
14 The Automated Analysis, Cataloging and Searching of Digital Image Libraries: A Machine Learning Approach	225
by Usama M. Fayyad and Padhraic Smyth	
14.1 Introduction	225
14.1.1 The Query Formulation Problem	226
14.1.2 Overview of Chapter	227
14.2 Case Study 1: The SKICAT System	227
14.2.1 Classifying Sky Objects	228
14.2.2 Classifying Faint Objects	229
14.2.3 Summary of Results	230
14.2.4 Summary of Benefits	230
14.3 Case Study 2: The Search for Volcanoes	231
14.3.1 The Approach	232
14.3.2 Current Status and Preliminary Results	233
14.4 Uncertainty in Ground Truth Labelling	233
14.4.1 Collecting Labelled Data	235
14.4.2 Modeling Multiple Expert Labellings	236
14.4.3 Experimental Results	238
14.4.4 Performance Evaluation using ROC Curves	238

14.5 Other Issues in Image Library Analysis	243
14.5.1 The Role of Prior Information	243
14.5.2 Modelling Spatial Context	245
14.5.3 Online Learning and Adaptation	245
14.5.4 Multi-Sensor and Derived Map Data	246
14.6 Discussion	247
14.7 Summary and Conclusion	248
V Prototypes/Applications	251
15 A Video Database System for Digital Libraries	253
by HongJiang Zhang, Stephen W. Smoliar, Jian Hua Wu, Chien Yong Low and Atreyi Kankanhalli	
15.1 Introduction	253
15.2 Video Parsing	254
15.3 Index Data Structure	256
15.4 Retrieval and Browsing	259
15.5 Summary	264
15.6 Acknowledgements	264
16 Developing The Scientific-Technical Digital Library at a National Laboratory	265
by Laurie E. Stackpole, Roderick D. Atkinson and John Yokley	
16.1 Introduction	265
16.2 The InfoNet	266
16.2.1 Impetus for Developing a Campus-wide Information System	266
16.2.2 Design and Development of the InfoNet	267
16.2.3 How the InfoNet Impacts Research Productivity	267
16.3 Research Reports Imaging System	268
16.3.1 Impetus for Optical Storage	268
16.3.2 Design Considerations for NRL's Prototype Imaging System	269
16.3.3 Design and Implementation of a Reports Catalog That Displays the Full Document	269
16.4 TORPEDO (The Optical Retrieval Project: Electronic Docu- ments Online)	270
16.4.1 Impetus for Delivery of Images to Researchers	270
16.4.2 Network Dissemination of Current Journals and Reports	271
16.4.3 Design and Development Considerations	271
16.4.4 Software Evaluation	272
16.4.5 Selected Software for Image Networking	276
16.5 Conclusion	278

17 DL-Raid: An Environment for Supporting Digital Library Services	281
Bharat Bhargava, Melliyal Annamalai, Shalab Goel, Shunge Li, Evaggelia Pitoura, Aidong Zhang and Yongguang Zhang	
17.1 Introduction	281
17.2 Digital Library Data Model	282
17.2.1 A Typology of Digital Library Data	283
17.2.2 Features of Object-Oriented Paradigm	284
17.2.3 Additional Features of Distributed Object-Oriented Systems	285
17.2.4 Digital Library Data Description - Layered Approach	285
17.3 DL-Raid System Model	287
17.3.1 The Architecture of DL-Raid	287
17.3.2 Extending O-Raid into DL-Raid	288
17.4 Partial Content-Based Retrievals in DL-Raid	290
17.4.1 Retrieval Methods	290
17.4.2 Supporting Image Classes	290
17.4.3 Probability-Based Fuzzy Retrievals	292
17.5 Communication Issues for Large Objects	292
17.5.1 WANCE Tool	293
17.5.2 Experimental Results	293
17.5.3 Effects on DL Data Retrieval Strategies	294
17.6 Conclusions and Future Work	295
Bibliography	301