LIST OF FIGURES

Fig. 1.1: General Architecture of a Web Search Engine 2
Fig. 1.2: Chapter-wise Organization of the Dissertation 10
Fig. 2.1: Internet Usage Statistics since 1995 14
Fig. 2.2: Growth curves for the Internet and the WWW 14
Fig. 2.3: Client-Server Model of Web Applications 16
Fig. 2.4: Types of Hyperlinks in an HTML Document 18
Fig. 2.5: Example Hyperlinked Structure depicting the Types of Hyperlinks 19
Fig. 2.6: The Basic Process of Information Retrieval 20
Fig. 2.7: Architecture of a Web Search Engine 26
Fig. 2.8: The Crawling Process of a Search Engine 27
Fig. 2.9: Algorithm for Crawler 28
Fig. 2.10: An Example stop-word list of twenty-five semantically nonselective words 30
Fig. 2.11: Example Suffix Tree for string BANANA 31
Fig. 2.12: The Concept of Dictionary and Postings 33
Fig. 2.13: Merging in Block Sort based Indexing 34
Fig. 2.14: Query Evaluation in a Typical Search Engine 35
Fig. 2.15: Distribution of Page Ranks 37
Fig. 2.16: Example Hyperlinked Structure 38
Fig. 2.17: Hubs and Authorities 42
Fig. 2.18: Algorithm to determine Base Set 43
Fig. 2.19: Algorithm to determine Hubs and Authorities 44
Fig. 2.20: An example of HITS operation 44
Fig. 2.21: A Query-Click Graph of Query Log 50
Fig. 2.22: A Query-Document Graph of Query Log 50
Fig. 2.23: A Query-Trace Graph of Query Log 50
Fig. 3.1: Taxonomy of Web Mining 55
Fig. 3.2: Representation of Documents in the Vector Space Model 57
Fig. 3.3: Graph Model of WWW 62
Fig. 3.4: Bow-Tie Structure of Web Graph 63
Fig. 3.5: (a) Page i and j co-cited by k, (b) Page i and j cite page k
Fig. 3.6: Web Usage Mining Process
Fig. 3.7: General Architecture of Web Usage Mining
Fig. 3.8: Schematic Structure of Session-Pageview Matrix
Fig. 3.9: The K-means Algorithm
Fig. 3.10: An Example Dendogram with 10 Objects (pages)
Fig. 3.11: The Hierarchical Agglomerative Clustering (HAC) Algorithm
Fig. 3.12: The Apriori Algorithm
Fig. 3.13: The GSP Algorithm
Fig. 3.14: Candidate Generation in GSP
Fig. 4.1: The P & P Annotation
Fig. 4.2: The P & P Framework of Web Search Engine
Fig. 4.3: Taxonomy of Proposed Pre and Post-Mining Techniques
Fig. 4.4: Knowledge Discovery via P & P Framework
Fig. 5.1: Refined Crawling and Indexing Process of a Search engine
Fig. 5.2: Data Structures for Crawling and Indexing Process
Fig. 5.3: Organization of Information in Proposed Index Structure
Fig. 5.4: A Sample Page R, and embedded Hyperlinks L
Fig. 5.5: Feedback Mechanism in a Search Engine
Fig. 5.6: Functioning of Crawl_Index Updater
Fig. 5.7: The Algorithm for Crawl_Index Updater
Fig. 5.8: Example Illustration of Crawl_Index Updater
Fig. 5.9: Approach taken by QUESEM
Fig. 5.10: The Expected output of QUESEM
Fig. 5.11: High-Level System Architecture of QUESEM
Fig. 5.12: Schema of Definition Repository
Fig. 5.13: Example Illustration of Definition Repository
Fig. 5.14: Definition Repository Generation Process
Fig. 5.15: Algorithm for Local Site Search for Query Terms
Fig. 5.16: Algorithm for Definition Generation/Annotation
Fig. 5.17: Definition Based Search Subsystem
Fig. 6.14: Proposed Format of Server Log Files 165
Fig. 6.15: Unequal Distribution of Link Weights 167
Fig. 6.16: Example Hyperlinked Structure with Link Visits 168
Fig. 6.17: Framework for Rank Optimization and Query Recommendation 171
Fig. 6.18: A Bipartite Graph of Query Log 173
Fig. 6.19: A Bipartite Graph for Query Log of Table 6.6 175
Fig. 6.20: Algorithm for GSP_QC 177
Fig. 6.21: Pictorial Representation of a Sequential Pattern 178
Fig. 7.1: Interface of QUESEM 187
Fig. 7.2 Result Screen of QUESEM for Query “apple” 188
Fig. 7.3: Result Pages after Browsing for “apple computer” 189
Fig. 7.4: Result Pages after Browsing for “apple mobile” 189
Fig. 7.5: Result Pages after Browsing for “apple fruit” 190
Fig. 7.6: Google’s Screen while submitting Query “apple” 190
Fig. 7.7: Google’s Results for Query “apple” 191
Fig. 7.8: Schema TERM for storing various Query Terms 191
Fig. 7.9: Schema STERM for storing Definitions of Query Terms 192
Fig. 7.10: State of TERM Relation 192
Fig. 7.11: State of STERM Relation 193
Fig. 7.12: Result Screen of QUESEM for Query “stand” 193
Fig. 7.13: Result Screen of QUESEM for Query “database” 194
Fig. 7.14: Result Screen of QUESEM for Query “search engine” 194
Fig. 7.15: Result Analysis of QUESEM for Definition “apple fruit” 195
Fig. 7.16: Result Analysis of QUESEM for Definition “apple computer” 195
Fig. 7.17: Interface for WSR 196
Fig. 7.18: Processing of User Query by WSR 197
Fig. 7.19: Page Similarity and Link Weight Calculation 198
Fig. 7.20: Rank Calculation for \( d=0.8 \) 198
Fig. 7.21: Final Ordered Results with respect to WSR 200
Fig. 7.22: Final Ordered Results with respect to Similarity Values 200
Fig. 7.23: Link Report Generated for the Web Graph 201
Fig. 7.24: WSR Processing for Query “JNU JNU Delhi” 201
Fig. 7.25: Final Ordered Results for Query “JNU JNU Delhi” 202
Fig. 7.26: Final Ordered Similarity Values for Query “JNU JNU Delhi” 202
Fig. 7.27: Variation of PR and WSR Values 203
Fig. 7.28: Variation of WSR with Similarity Values 204
Fig. 7.29: Variation of PR with Similarity Values 204
Fig. 7.30: State1 of Link_Log 206
Fig. 7.31: PRLV Values corresponding to State1 207
Fig. 7.32: State2 of Link_Log 207
Fig. 7.33: PRLV Values corresponding to State2 207
Fig. 7.34: State3 of Link_Log 208
Fig. 7.35: PRLV Values corresponding to State3 208
Fig. 7.36: Variation of PRLV with PR and WSR 209
Fig. 7.37: Interface of Query Log Updation 210
Fig. 7.38: Finding Similarity based on Query Contents 210
Fig. 7.39: Finding Similarity based on Clicked URLs 211
Fig. 7.40: Cluster Generation based on the Combined Measure 211
Fig. 7.41: Generated Favored Queries 212
Fig. 7.42: Search Results for Query “data mining and warehousing” 213
Fig. 7.43: Previous Rank of URL 213
Fig. 7.44: Updated Rank of URL 214
Fig. A.1: Result Screen of Google for Query “data mining” 214
Fig. A.2: Similarity Variations of First 25 Pages for Query “data mining” 216
Fig. A.3: Similarity Variations of First 25 Pages for Query “data warehousing” 218
Fig. B.1: The Web Graph for Analysis 221
Fig. B.2: Web Page a i.e. university.html 222
Fig. B.3: Web Page b i.e. university of Delhi.html 223
Fig. B.4: Web Page c i.e. university of Delhi.html 223
Fig. B.5: Web Page d i.e. anna university.html 224
Fig. B.6: Web Page e i.e. university of hyderabad.html 224
Fig. B.7: Web Page f i.e. jnu.html 225
Fig. B.8: Web Page g i.e. IITT Hyderabad.html 225
Fig. B.9: Web Page h i.e. IIT Delhi.html 226
Fig. B.10: Web Page i i.e. DU project.html 226
Fig. B.11: Web Page j i.e. Vaibhav V. Kaware pune.html 227
Fig. B.12: Web Page k i.e. anna university project.html 227
Fig. B.13: Web Page l i.e. XYZ Pvt Ltd.html 228
Fig. B.14: Web Page m i.e. Project in JNU.html 228
Fig. B.15: Web Page n i.e. R & D of IITT Hyderabad.html 229
Fig. B.16: Web Page o i.e. Research Areas IIT Delhi.html 229
Fig. B.17: Web Page p i.e. Sorav Bansal IIT Delhi.html 230
Fig. B.18: Web Page q i.e. Prof. K. K. Biswas.html 230
Fig. B.19: Web Page r i.e. Naveen Garg IIT DELHI.html 231
Fig. B.20: Web Page s i.e. Sanjiva Prasad IIT Delhi.html 231
Fig. B.21: Web Page t i.e. Anshul Kumar IIT Delhi.html 232
Fig. B.22: Web Page u i.e. Rakesh Kumar.html 232
Fig. B.23: Web Page v i.e. K.G. Saxena JNU.html 233
Fig. B.24: Web Page w i.e. V. Subramanian jnu.html 233
Fig. B.25: Web Page x i.e. Arun K. Attri jnu.html 234
Fig. B.26: Web Page y i.e. V. Rajamani jnu.html 234
Fig. B.27: Web Page z i.e. Indian Council of Agricultural Research.html 235