# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Particulars</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>viii-ix</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>xi-xiii</td>
</tr>
<tr>
<td>1.</td>
<td><strong>INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Background and Motivation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Hypertext and the Web</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.3 Search Engines</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1.4 Web Crawlers</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.4.1 Parallel Web Crawler</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1.5 Objectives of the Proposed Research Work and Research Questions</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1.6 Main Contribution of the Thesis</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1.7 Structure of the Thesis</td>
<td>13</td>
</tr>
<tr>
<td>2.</td>
<td><strong>LITERATURE REVIEW</strong></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>2.1 Introduction to Hypertext system</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>2.1.1 Information Retrieval</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2.1.2 Combining Information Retrieval and Hypertext</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>2.1.3 Internet Resource Discovery</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>2.1.4 Web Search Engines</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>2.1.5 Meta-Searching</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>2.2 Distributed Systems</td>
<td>19</td>
</tr>
</tbody>
</table>
2.2.1 Architecture of Distributed Systems 19
2.2.2 Load Balancing 19
2.2.3 Resource Location 20

2.3 Web Characterization 20
2.3.1 Methods for Sampling 20
2.3.2 Web Dynamics 20
   2.3.2.1 Estimating Freshness and Age 21
   2.3.2.2 Characterization of Web Page Changes 21
2.3.3 Link Structure 21
   2.3.3.1 Scale-Free Networks 22
   2.3.3.2 Macroscopic Structure 22
2.3.4 User Sessions On The Web 22

2.4 Indexing and Querying Web pages 23
2.4.1 Inverted Index 23
2.4.2 Distributing Query Load 24
2.4.3 Text-Based Ranking 24

2.5 Connectivity-Based Ranking 24
2.5.1 Query-Independent Ranking 25
2.5.2 Query-Dependent Ranking 25

2.6 Web Crawling Issues 26
2.6.1 Selection Policy 26
2.6.2 Re-Visit Policy 28
2.6.3 Politeness Policy 29
2.6.4 Parallelization Policy 29

2.7 Examples of Web Crawlers 29

2.8 Web Caching 31
2.8.1 Web Indexers 32
2.8.2 Search Ranking and Community Discovery Systems 32
2.8.3 Meta-Search Engines 32
2.8.4 Focused Crawlers 33

2.9 Crawling Strategies 33
## 2.9.1 Crawling Algorithm

- **2.9.1.1 Fish Algorithm** 35
- **2.9.1.2 Shark Algorithm** 36
- **2.9.1.3 Breadth First Search Algorithm** 38
- **2.9.1.4 Depth First Search Algorithm** 39
- **2.9.1.5 Best First Search Algorithm** 39

### 2.10 Other Related Topics

- **2.10.1 Monitoring Technique** 39
- **2.10.2 Site Selection** 40
- **2.10.3 Frequency of Page Change** 40
- **2.10.4 Lifespan of a Web Page** 40
- **2.10.5 Web Page Change** 41

### 2.11 Conclusion

## 3. ARCHITECTURE OF MIGRATING PARALLEL WEB CRAWLER

- **3.1 Introduction** 42
- **3.2 Problem Definition & Scope** 43
- **3.3 Proposed Architecture of Migrating Parallel Web Crawler** 44
- **3.4 Algorithm** 45
- **3.5 Brief Description of Modules** 46
- **3.6 Conclusion** 47

## 4. DOMAIN SPECIFIC AND INCREMENTAL CRAWLING

- **4.1 Introduction** 48
- **4.2 Problems in Generic Web Crawlers** 50
- **4.3 Issues in Migrating Parallel Web Crawlers** 50
  - **4.3.1 Collection Updating in Place** 52
  - **4.3.2 Page Refresh Frequency** 53
4.4 Coordination in Migrating Parallel Web Crawler 53
4.5 Crawling Modes for Static Assignment 54
4.6 Evaluation Model in Migrating Parallel Web Crawler 55
4.7 Incremental Crawling 57
4.8 Domain Specific and Incremental Crawling in Migrating Parallel Web Crawler
   4.8.1 Central Coordinator System 62
   4.8.2 Crawling Process 63
4.9 Algorithm 64
4.10 Crawler Implementation 65
   4.10.1 java.lang Package 65
      4.10.1.1 Runnable Interface 66
         4.10.1.1.1 Timer Class 66
         4.10.1.1.1 Chronicle Class 66
         4.10.1.1.2 Crawler Class 67
      4.10.1.2 Cloneable Interface and clone () 69
         4.10.1.2.1 Download Parameter Class 69
   4.10.2 HTMLTransformer Class 69
      4.10.2.1 LinkTransformer Class 70
         4.10.2.1.1 RewritableLinkTransformer Class 70
         4.10.2.1.1.1 Concatenator Class 71
         4.10.2.1.1.2 RecordTransformer Class 71
      4.10.2.2 CrawlListener Interface 71
         4.10.2.2.1 EventLog Class 72
   4.10.3 java.io.Serializable 73
      4.10.3.1 Action Interface 74
      4.10.3.2 PagePredicate Interface 74
5. CHANGE DETECTION IN MIGRATING PARALLEL WEB CRAWLER: A NEURAL NETWORK BASED APPROACH

5.1 Introduction
5.2 Change Detection Algorithms
  5.2.1 Types of Change
  5.2.2 Detecting Changes
5.3 Background on Neural Networks
5.4 Neural Network Based Model For Change Detection Of Web Pages
  5.4.1 Neuron Model
  5.4.2 Network Architecture
5.5 Neural Network Based Change Detection Of Web Pages
5.6 Description
  5.6.1 Method Followed For Detecting Change In The Structure
  5.6.2 Method Followed For Detecting Change In The Content
5.7 Conclusion

6. SOLUTION TO NETWORK TRAFFIC PROBLEM IN MIGRATING PARALLEL CRAWLERS USING FUZZY LOGIC

6.1 Introduction
6.2 Quality and Network Metrics
   6.2.1 Geographic Distance
   6.2.2 Latency
   6.2.3 Correlation between Metrics
6.3 Case Study of Crawler Load
6.4 Fuzzy Inference Systems and Fuzzy Logic
6.5 Proposed Solution
6.6 Description
6.7 Result
6.8 Conclusion

7. VALIDATION OF ARCHITECTURE OF MIGRATING PARALLEL WEB CRAWLER USING FINITE STATE MACHINE

7.1 Introduction to Finite-State Machine
7.2 Validation of the Proposed Architecture
7.3 Steps for the Validation of Architecture using FSM
7.4 Validation Of Migrating Parallel Web Crawler Using Finite State Machine
7.5 Empirical Validation and Statistical Analysis
7.6 Result
7.7 Conclusion

8. CONCLUSION AND FUTURE WORK
8.1 Conclusion 148
8.2 Future Work 150

LIST OF PUBLICATIONS 151

REFERENCES 152