

## List of Figures

Fig. 1.1: Cyclic architecture for search engines	2
Fig. 1.2: Simple query interfaces from book domain	7
Fig. 1.3: Main topics covered in this thesis	9
Fig. 2.1: Elements of a Search Engine	17
Fig. 2.2: Architecture of a Typical Crawler	20
Fig. 2.3: Algorithm for Crawler	20
Fig. 2.4: General Architecture of a Parallel Crawler	24
Fig. 2.5: The Mapping Process	26
Fig. 2.6: The Crawl Manager	28
Fig. 2.7: The Crawl worker	29
Fig. 2.8: Architecture of a Focused Crawler	31
Fig. 2.9: Sample Augmented XML code	33
Fig. 2.10: TOC file for sample code of Fig. 2.9	33
Fig. 2.11: Modified Category tree for Search Engine	35
Fig. 2.12: High-Level View of Context Driven Focused Crawler	35
Fig. 2.13: Relationship between components of CDFC	37
Fig. 2.14: General Architecture of a Mobile crawling system.	39
Fig. 2.15: Architecture of Incremental Crawler	43
Fig. 2.16: Form Focused Crawler	44
Fig. 2.17: Small Distributed Crawler configuration	46
Fig. 3.1: Query Interfaces from amazon.com and makemytrip.com	49
Fig. 3.2: Distribution of Hidden Web Sites by Content Type	51
Fig. 3.3: Ten year growth trends in information contents	52
Fig. 3.4: User-Search Interface Interaction	54
Fig. 3.5: Execution Sequence of Deep Web Crawler	55
Fig. 3.6: Architecture of Deep Web Crawler	56
Fig 3.7: Architecture of Hidden Web Crawler	60
Fig 3.8: Classes of HTML forms	63
Fig. 4.1: Crawler Search Interface Interaction	66

Fig. 4.2: Architecture of a Domain-specific Hidden Web Crawler	67
Fig. 4.3: Proposed Architecture for Search Interface Crawling using Augmented hypertext	69
Fig. 4.4: Algorithm for URL Dispatcher	70
Fig. 4.5: Domain hierarchy in URL Database	71
Fig. 4.6: Algorithm for Form Identifier	72
Fig. 4.7: Algorithm for Downloader	73
Fig. 4.8: The Domain-specific Interface Mapper	76
Fig. 4.9: Hierarchical representation of a query interface in a particular domain	77
Fig. 4.10: Two query interfaces in book domain	77
Fig. 4.11: Hierarchical representation of two query interfaces in book domain	77
Fig. 4.12: Example of two Similarity Value Matrices	81
Fig. 4.13: Detailed Schematic Diagram of SVM Generator and Selector module	82
Fig. 4.14: Example of matching process	83
Fig. 4.15: Algorithm for SVM Generator and Selector	84
Fig. 4.16: Structure of Mapping Knowledge Base	84
Fig. 4.17: Merging algorithm to merge two query interfaces $S_i$ and $S_j$ in same domain	87
Fig. 4.18: The Unified Search Interface (USI) for the book domain	90
Fig. 4.19 (a): Automatic form filling process	89
Fig. 4.19 (b): Data Extraction Engine	89
Fig. 4.20: Result snippets from Google	90
Fig. 4.21: Architecture of AKSHR	92
Fig. 5.1: Home Page of DSHWC	96
Fig. 5.2: Interface 1 for Book Domain	98
Fig. 5.3: Search Interface 2 for Book Domain	99
Fig. 5.4: Search Interface 3 for Book Domain	99
Fig. 5.5: Search Interface 4 for Book Domain	100
Fig. 5.6: Search Interface 5 for Book Domain	100
Fig. 5.7: No. of comparisons Vs Mapping in Knowledgebase at Threshold=0.60	102
Fig. 5.8: Mapping Results at Threshold=0.60	102

Fig. 5.9: Mapping Results at Threshold=0.65	103
Fig. 5.10: Combined Mapping Results at Threshold=0.60	104
Fig. 5.11: Combined Mapping Results at Threshold=0.65	104
Fig. 5.12: Unified Search Interface for Books domain	105
Fig. 5.13: Unified Search Interface (USI) for Books domain	106
Fig. 5.14: Merging results of book domain at threshold=0.60	107
Fig. 5.15: Merging results of book domain at threshold=0.65	107
Fig. 5.16: Merging results at threshold=0.60	108
Fig. 5.17: Average merging results at threshold=0.65	109
Fig. 5.18: Snapshot of Websites	110
Fig. 5.19: Result Page On Amazon.Com.	110
Fig. 5.20: Result Page On Books.Com.	111
Fig. 5.21: Result Page On Rediffbooks.Com.	111
Fig. 5.22: Result Page On Magabalabooks.Com.	112
Fig. 5.23: Result Page On Books.Com.	112
Fig. 5.24: Result Page On Authorhouse.Com.	113
Fig. 5.25: Results of offline submission	113
Fig. 5.26: Results of offline submission	114
Fig. 5.27: Precision, Recall and F-measure for Books Domain.	115
Fig. 5.28: Precision, Recall and F-measure for Airlines Domain.	115
Fig. 5.29: Precision, Recall and F-measure for Electronics Domain.	116
Fig. 5.30: Precision, Recall and F-measure for Automobiles Domain.	117