

ABSTRACT

The World Wide Web (WWW), the largest and most frequently accessed public repository of information ever developed, contains large number of web pages interconnected through hyperlinks. It has grown from a few thousand pages in 1993 to more than a couple of billion pages at present. It is easily accessible through internet connection, though it is often more difficult to locate the resources that are relevant to the given need. Due to this problem of *information overkill*, the development of newer and powerful information retrieval tools (search engines) assume importance.

A Web search engine searches for information consisting of web pages, images and other types of files lying on the WWW. In order to maintain index at the search engine's side, a search engine, generally, employs a web crawler to download web pages.

c

The crawler, on its turn, is given a starting set of URLs corresponding to which pages are downloaded from the web. Before storing the web pages, the crawlers extract URLs appearing in the retrieved pages and pass them to a crawler control module which determines what links to visit next and gives the selected links back to the crawler and so on. The downloaded pages are stored into a page repository.

The Web can be divided into two parts: Surface Web and Deep Web. The Surface Web refers to the static Web pages that can be crawled and indexed by popular search engines, also termed as Publically Indexable Web (PIW). On the other hand, the Deep Web refers to the contents stored in Web databases and published by dynamic Web pages wherein people access web databases through specified query interfaces.

Infact, there are more than 300,000 Deep Web databases and 450,000 query interfaces available in the hidden web and the two figures are still increasing quickly. Besides the scale of Web databases, the contents in Web databases span well across all topics ranging from agriculture to nuclear domain. Some Deep Web portal services provide Deep Web directories that classify Web databases in some taxonomies, contain large amount of high

quality information. However, these sites hidden behind search interfaces can not be crawled by traditional crawlers. Infact, crawling hidden Web is a very challenging problem especially because of following two fundamental reasons:

- Access to these databases is provided only through restricted search interfaces, intended to be filled manually.
- Besides the access through search interfaces, the sheer size of the hidden web is too large i.e. about 400 to 500 times larger than the size of the *Surface Web*. As a result, it is not prudent to attempt comprehensive coverage of the hidden Web and therefore, there is need to develop a domain-specific crawler for hidden Web.

In this thesis, design and development of a novel framework for an Extensible and Scalable Domain-Specific Hidden Web Crawler (DSHWC) is being reported. It is not only capable of crawling the hidden web but can also efficiently deal with databases hidden behind search interfaces containing single and multiple attributes as well.

CONTENTS

Certificate	i
Candidate's Declaration	ii
Acknowledgements	iii
Dedication	iv
Abstract	v
Contents	vii
List of Figures	xi
List of Tables	xiv
CHAPTER 1: INTRODUCTION	1
1.1 GENERAL	1
1.2 WEB SEARCHING	2
1.3 MOTIVATION	3
1.4 HIDDEN WEB (Deep or Invisible Web)	5
1.5 CHALLENGES OF HIDDEN WEB CRAWLING	6
1.6 ORGANIZATION OF THESIS	8
CHAPTER 2: SEARCH ENGINES & CRAWLERS: A REVIEW	11
2.1 GENERAL	11
2.2 SEARCH ENGINES	<u>12</u>
2.2.1 BOOLEAN MODEL	13
2.2.2 VECTOR SPACE MODEL	14
2.2.3 PROBABILISTIC MODEL	15
2.2.4 ELEMENTS OF A WEB SEARCH ENGINE	16
2.2.5 TYPE OF DATA RETRIEVED BY SEARCH ENGINE	18
2.3 THE CRAWLERS	19

2.3.1	ROBOT.TXT: A STANDARD FOR ROBOT EXCLUSION	21
2.3.1.1	User Agent	22
2.3.1.2	Disallow	22
2.3.2	TYPES OF CRAWLERS	23
2.3.2.1	Parallel Crawler	23
2.3.2.2	Parallel Crawler using Augmented Hypertext Documents (PARCAHYD)	25
2.3.2.3	Focused Crawler	30
2.3.2.4	Context Driven Focused Crawler (CDFC)	32
2.3.2.5	Mobile or Migrating Crawlers	38
2.3.2.6	Incremental Crawler	40
2.3.2.7	Form Focused Crawler	44
2.3.2.8	Distributed crawler	45
CHAPTER 3: HIDDEN WEB CRAWLING STRATEGIES: A REVIEW		49
3.1	HIDDEN WEB	49
3.1.1	BROAD, RELEVANT COVERAGE	50
3.1.2	HIGHER QUALITY	50
3.1.3	ORIGINAL HIDDEN CONTENT EXCEEDS ALL PRINTED GLOBAL CONTENT	52
3.2	HIDDEN WEB CRAWLING TECHNIQUES	54
3.3.1	DEEP WEB CRAWLER	54
3.2.1.1	Architecture of Deep Web Crawler	56
3.2.1.2	LVS Table	58
3.2.2	CRAWLING FOR DOMAIN-SPECIFIC HIDDEN WEB RESOURCES	60
3.2.2.1	The Local Crawler	61
3.2.2.2	The Form analyzer	62
3.2.2.3	The Query Prober	63

CHAPTER 4: AN EXTENSIBLE AND SCALABLE TASK-ORIENTED (DOMAIN-SPECIFIC) HIDDEN WEB CRAWLER (DSHWC) 66

4.1	INTRODUCTION	66
4.2	PHASE I: SEARCH INTERFACE CRAWLING	68
4.2.1	URL DISPATCHER	69
4.2.2	URL BUFFER	70
4.2.3	LINK DATABASE	70
4.2.4	LOS BUFFER	71
4.2.5	FORM IDENTIFIER	71
4.2.6	SEARCH INTERFACE REPOSITORY	72
4.2.7	DOWNLOADER	73
4.3	PHASE II: DOMAIN-SPECIFIC INTERFACE MAPPING (<i>DSIM</i>)	74
4.3.1	PARSING	75
4.3.2	SEMANTIC MATCHING	78
4.3.2.1	Fuzzy Matching	78
4.3.2.2	Domain Specific Thesaurus	79
4.3.2.3	Data Type Matching	80
4.3.3	MAPPING GENERATION	82
4.3.4	MAPPING KNOWLEDGEBASE	83
4.4	PHASE III: MERGING THE QUERY INTERFACES	85
4.5	PHASE IV: AUTOMATIC FORM FILLING	88
4.6	PHASE V: RESPONSE PAGE ANALYSIS	90
4.7	AKSHR: AN ALTERNATE FRAMEWORK FOR DOMAIN-SPECIFIC HIDDEN WEB CRAWLER	91

CHAPTER 5: IMPLEMENTATION & RESULT ANALYSIS OF DSHWC 95

5.1	GENERAL	95
5.2	PERFORMANCE METRICS	96
5.2.1	DATA SETS	97
5.2.2	EXPERIMENTS	97

5.3	PHASE I: SEARCH INTERFACE CRAWLING	98
5.4	PHASE II: DOMAIN-SPECIFIC INTERFACE MAPPING (DSIM)	101
5.5	PHASE III: SEARCH INTERFACE MERGING	105
5.6	PHASE IV: FORM FILLING	109
5.7	OVERALL ANALYSIS OF DSHWC	114
	5.7.1 BOOKS DOMAIN	114
	5.7.2 AIRLINES DOMAIN	115
	5.7.3 ELECTRONICS DOMAIN	116
	5.7.4 AUTOMOBILES DOMAIN	116
	5.7.5 COMPARISON OF ALL DOMAINS	117
5.8	COMPARISON OF DSHWC AND AKSHR WITH EXISTING HIDDEN WEB CRAWLER	119
CHAPTER 6: CONCLUSION & FUTURE SCOPE		120
6.1	CONCLUSION	120
6.2	FUTURE SCOPE	121
REFERENCES		123
APPENDIX A		133
APPENDIX B		140