

# **Refreshing Search Engine Index by Swarm Intelligence over Social Networks**

Kesara Nanayakkara Rathnayake

09/10010



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Dissertation submitted to the Faculty of Information Technology,  
University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of  
the Degree of MSc in Artificial Intelligence.

August 2011

## Declaration

I declare that this dissertation does not incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organization.

Kesara Nanayakkara Rathnayake

Name of Student

Signature of Student

Date



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Supervised by

Prof. Asoka Karunananda

Name of Supervisor(s)

Signature of Supervisor(s)

Date

## **Dedication**

This project is dedicated to my parents who gave me all the support and love. Without them completion of this work would not have been possible.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## **Acknowledgements**

I wish to thank my supervisor Prof. Asoka Karunananda for his support and guidance. I wouldn't have finished this work without my parents' and fiancées' support, love and encouragement. I also wish to thank all my friends in MSc AI batch 2 and my colleagues at office for their support and encouragement.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Abstract

Humans make lot of decisions in their day-to-day life. In order to make right decisions they need more information. The WWW contains enormous amount of information. It is a huge complex system. Finding correct up-to-date information on WWW is a difficult task. Search engines make that task easier. Search engines are the main tool used to search WWW. If a search engine starts searching for a web page as soon as user enters the query, the searching will take almost infinite time since WWW is a vast collection of web pages. The reason search engine provides information so quickly is due to the fact that search engine has already crawled the WWW and stored data in an index. Index is one the most important components of the search engine.

Web crawlers are used to populate index of a search engine by crawling web sites. Search engine without an up-to-date index is pointless since web pages in WWW gets updated all the time and more and more web pages and web sites emerges. Therefor this index has to be updated regularly. Maintaining an index with up-to-date information on such a complex system is a difficult task. This project addresses the issue of inefficient information retrieval in search engine domain.

The social networks and other social media reflect current world trends. Therefor social networks can be used to identify current world trends. This project uses swarm intelligence to identify current world trends via social networks. This is done by collecting and analyzing status messages and micro-blog messages that users publish on popular social networks. Swarm intelligence is used to analyze the status messages and micro-blog messages on social networks; and identifies the current world trends.

A MAS based web crawler system was designed and developed to crawl the WWW based on current world trends identified by swarm intelligence based analysis of status messages and micro-blog messages on popular social networks. The proposed MAS based crawler system was compared against a conventional crawler system on identifying newly updated web pages. The proposed MAS based web crawler system, crawls web efficiently to retrieve updated web pages based on their utility value or the importance on timely manner based on the identified current world trends.

# Contents

	<b>Page</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Introduction	1
1.2 Aim	3
1.3 Objectives	3
1.4 Solution	3
1.5 Structure of the Report	4
1.6 Summary	4
<b>Chapter 2 Current Issues in Search Engine Domain</b>	<b>5</b>
2.1 Introduction	5
2.2 Related Work	5
2.2.1 Crawlers based on Statistical Analysis	5
2.2.2 Incremental Crawlers	5
2.2.3 Poisson Process based Crawlers	6
2.2.4 Genetic Algorithms based Crawlers	7
2.2.5 MAS based Crawlers	7
2.2.6 Identifying Content Changes	8
2.2.7 Duplicate Detection	8
2.2.8 Identifying Web Spam	9
2.3 Problems Identified	9
2.4 Summary	13
<b>Chapter 3 Multi Agent Systems and Swarm Intelligence</b>	<b>14</b>
3.1 Introduction	14
3.2 Multi Agent Systems	14
3.2.1 Utility based Agents	15
3.3 Swarm Intelligence	16
3.4 Summary	17
<b>Chapter 4 MAS based Approach to Efficient Data Retrieval</b>	<b>18</b>
4.1 Introduction	18
4.2 Inputs	18
4.3 Outputs	19
4.4 Process	19

4.5 Users	19
4.6 Features	20
4.7 Summary	21
<b>Chapter 5 MAS based Crawler System Design</b>	<b>22</b>
5.1 Introduction	22
5.2 Design of the MAS based Web Crawler System	22
5.2.1 Response Agents in the System	22
5.2.2 Request Agents in the System	24
5.2.3 Social Network Spy Agents	24
5.2.4 World Trend Agent	24
5.2.5 Seed List Agent	25
5.2.6 Indexer Agent	25
5.2.7 Crawler Agents	25
5.3 Search Engine	26
5.3.1 Crawler System	26
5.3.2 Indexer and Query Engine	27
5.3.3 Search Engine Frontend	27
5.4 Life of a Web Page	28
5.5 Life of a Status Message	29
5.6 Life of a Query	30
5.7 Summary	30
<b>Chapter 6 Implementation of the MAS Based Crawler System</b>	<b>31</b>
6.1 Introduction	31
6.2 MAS based Crawler System	31
6.2.1 Crawler Agent	31
6.2.2 Social Network Spy Agent	32
6.2.3 World Trend Identification Agent	33
6.2.4 Seed List Agent	35
6.2.5 Indexer Agent	35
6.3 Search Engine Implementation	36
6.3.1 Data Retrieval System	36
6.3.2 Indexer and Query Engine	36
6.3.3 Search Engine Frontend	37

6.4 Connected Social Networks	38
6.4.1 Twitter	38
6.4.2 Virtual Social Network	38
6.5 Current World Trend Identification Process	39
6.6 Summary	40
<b>Chapter 7 Evaluation</b>	<b>41</b>
7.1 Introduction	41
7.2 Evaluation Strategy	41
7.2.1 Identification of Updated Web Sites	41
7.2.2 Efficiency in Crawling	42
7.3 Experiment Setup	42
7.3.1 Identification of Updated Web Sites	42
7.3.2 Efficiency in Crawling	43
7.4 Results of the Experiment	43
7.4.1 Results of Identification of Updated Web Sites	43
7.4.2 Efficiency in Crawling	45
7.5 Summary	46
<b>Chapter 8 Conclusion and Further Works</b>	<b>47</b>
8.1 Introduction	47
8.2 Conclusion	47
8.3 Further Works	48
8.4 Summary	49
<b>References</b>	<b>50</b>
<b>Appendix A Architecture of a Search Engine</b>	<b>53</b>
<b>Appendix B UML Diagrams</b>	<b>54</b>
<b>Appendix C Screenshots</b>	<b>58</b>
<b>Appendix D Sample Seed List Ontology</b>	<b>62</b>
<b>Appendix E Sample Topic-Words Mapping Ontology</b>	<b>63</b>
<b>Appendix F Test Results for Identification of Updated of Web Pages</b>	<b>64</b>
<b>Appendix G Test Results for Efficiency of Web Crawlers</b>	<b>66</b>



# List of Figures

	<b>Page</b>
Figure 3.1 – Architecture of Common MAS	15
Figure 3.2 – Simple MAS Setup for SI based Classification	17
Figure 4.1 – Input, process, output of the MAS based crawler system	18
Figure 5.1 – Top Level Architecture	23
Figure 5.2 – FMS of the Proposed System	24
Figure 5.3 – System Wide Component Diagram of the Search Engine	26
Figure 5.4 – Life of a Web Page	28
Figure 5.5 – Life of a Status Message	29
Figure 5.6 – Life of a Search Query	30
Figure 7.1 – Conventional Web Crawlers – Discovery Results (cumulated)	43
Figure 7.2 – MAS based Web Crawlers – Discovery Results (cumulated)	44
Figure 7.3 – Cumulative Success Rate	44
Figure 7.4 – Efficiency in Crawling for MAS based Web Crawler System	45
Figure 7.5 – Efficiency in Crawling for Conventional Web Crawler System	45
Figure A.1 – Architecture of a Search Engine	53
Figure B.1 – Class Diagram of the MAS based Crawler System	54
Figure B.2 – Activity Diagram of Spy Agent	55
Figure B.3 – Activity Diagram of Crawler Agent	56
Figure B.4 – Activity Diagram of World Trend Identification Agent	57
Figure C.1 – Screenshot of Crawler System Console	58
Figure C.2 – Unfiltered World Trend Tag Cloud	59
Figure C.3 – Filtered World Trend Tag Cloud	59
Figure C.4 – World Trend Evolution	59
Figure C.5 – Screenshot of Search Engine Frontend	60
Figure C.6 – Screenshot of Virtual Social Network	61

## List of Tables

	<b>Page</b>
Table 2.1 – Issues with Various Approaches	12
Table 6.1 – Topics and Related Words	40
Table 6.2 – Sample Weights	40
Table F.1 – Web Page Discovery Test	65
Table F.2 – Success Rates	65
Table G.1 – Number of Web Crawls within One Hour	66



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)