References

[1] Abraham A., Das S. and Roy S. (2007), *Swarm Intelligence Algorithms for Data Clustering*, Soft Computing for Knowledge Discovery and Data Mining Book, Part IV, pp 279-313

[2] Ali H. (2008), *Self Ranking and Evaluation Approach for Focused Crawler Based on Multi-Agent System*, The International Arab Journal of Information Technology, 5(2), pp 183-191

[3] Artail H. and Fawaz K. (2008), *A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations*, Data & Knowledge Engineering 66(2008), pp 326-337

[4] Batsakis S., Petrakis E. and Milios E. (2009), *Improving the performance of focused web crawlers*, Data & Knowledge Engineering 68(2009), pp 1001-1013

[5] Brewington B. E. and Cybenko G. (2000), *How dynamic is the Web?*, Computer Networks 33(1–6), pp 257-276

[6] Broder A. and et al. (1997), *Syntactic clustering of the web*, In proceedings of the 6th International World Wide Web Conference, Santa Clara, USA

[7] Chauhan N. and Sharma A. (2007), *Design of an Agent Based Context Driven Focused Crawler*, BVICAM's International Journal of Information Technology, 1(1), pp 61-66

[8] Cho J. and Garcia-Molina H. (2000), *The evolution of the web and implications for an incremental crawler*, In proceedings of the 26th International Conference on Very Large Data Bases, pp 200-209, San Francisco, USA

[9] Cho J. and Garcia-Molina H. (2003), *Estimating frequency of change*, ACM Transactions on Internet Technology (TOIT) 3(3), pp 256-290

[10] Fetterly D. and et al. (2003), *A Large-Scale Study of the Evolution of Web Pages*, Journal of Software Practice and Experience 34(2), pp 213-237

[11] Gloor P. and et al. (2009), *Web Science 2.0: Identifying Trends through Semantic Social Network Analysis*, In proceedings of IEEE Conference on Social Computing (SocialCom-09), Aug 29-31, Vancouver, Canada

[12] Gottron T. (2009), *Detecting Website Redesigns via Template Similarity on Streams of Documents*, In proceedings of 3rd International Conference of Internet Technologies and Applications (ITA09), Wrexham, UK

[13] Grimes C. and Ford D. (2008), *Estimation of Web Page Change Rates*, In proceedings of the Joint Statistical Meetings, Denver, USA

[14] Grimes C., Ford D. and Tassone E. (2008), *Keeping a search engine index fresh: risk and optimality in estimating refresh rates*, In proceedings of INTERFACE 2008: Culture & Technology, Ottawa, Canada

[15] Johnson J., Tsioutsiouliklis K. and Giles C. L. (2003), *Evolving strategies for focused web crawling*, In proceedings of the 20th International Conference on Machine Learning (ICML 2003), Washington DC, USA

[16] Kobayashi M. and Takeda K. (2000), *Information retrieval on the web*, ACM Computing Surveys (ACM Press) 32(2), pp 144-173

[17] Lawrence S. and Giles C. L. (1999), *Accessibility of information on the web*, Nature, 400, pp 107-109

[18] Matloff N. (2005), *Estimation of internet file access modification rates from indirect data*, ACM Transactions on Modeling and Computer Simulation 15, pp 233–253

[19] Menczer F. and Belew R. K. (2000), *Adaptive retrieval agents: Internalizing local context and scaling up to the web*, Machine Learning, 39(1), pp 203-242

[20] Olston C. and Najork M. (2010), *Web Crawling*, Foundations and Trends in Information Retrieval 4(3), pp 175-246

[21] Palathingal P., Potok T. and Patton R. (2005), *Agent Based Approach for Searching, Mining and Managing Enormous Amounts of Spatial Image Data*, In proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, pp 351-356, Florida, USA

[22] Panait L. and Luke S. (2005), *Cooperative multi-agent learning: The state of the art*, Autonomous Agents and Multi-Agent Systems, 11(3), pp 387-434

[23] Vizine1 A. L., de Castro1 L. N. and Gudwin R. R., *Text Document Classification Using Swarm Intelligence*, In proceedings of the 2005 IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS'05), pp 134-139, Waltham, USA

Appendix A

Architecture of a Search Engine

Figure A.1 illustrates the architecture of a common search engine. A common search engine consists of components like crawlers, document processors, indexer, query engine etc.

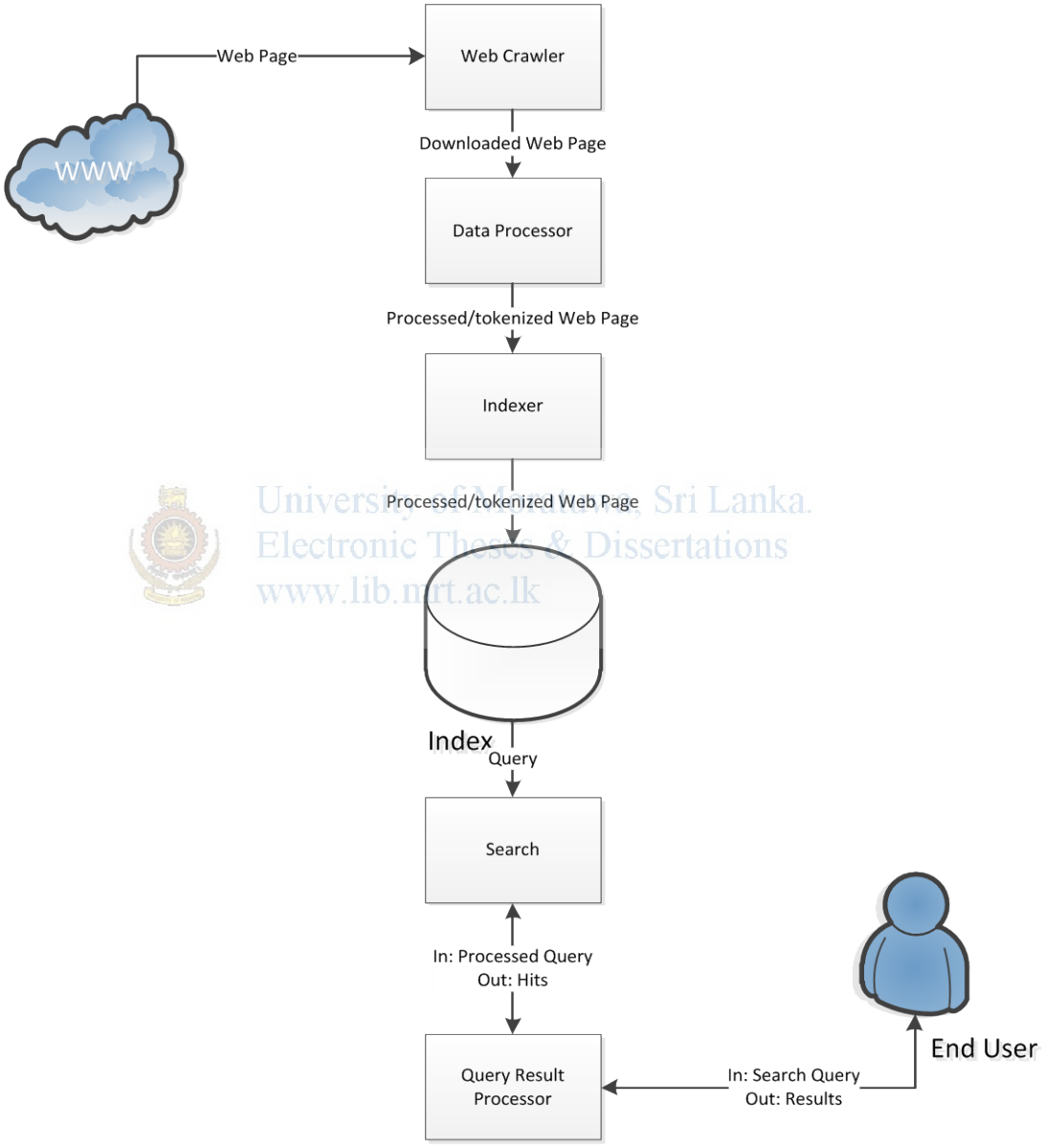

Figure A.1: Architecture of a Search Engine

Appendix B

UML Diagrams

This section includes UML diagrams of various components of the MAS based crawler system. Figure B.1 illustrates the class diagram of the MAS based crawler system. Section 6.2 discusses about the MAS based crawler system and it's classes.

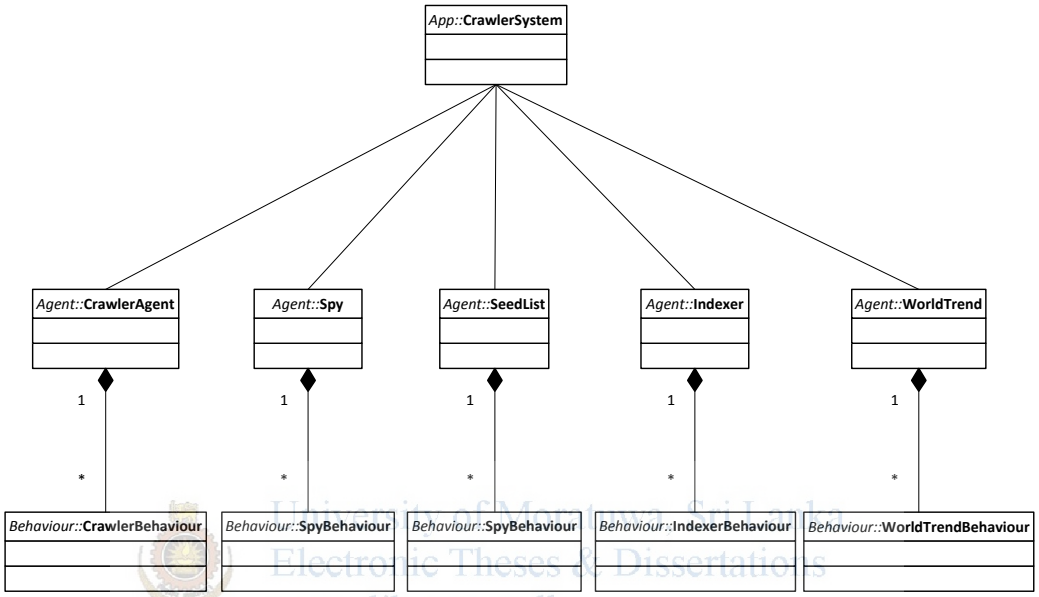

Figure B.1: Class Diagram of the MAS based Crawler System

Figure B.2 illustrates activities of the spy agent in the form of activity diagram. The spy agent is used to retrieve status messages from the social networks. Section 6.2.2 discusses more about spy agents.

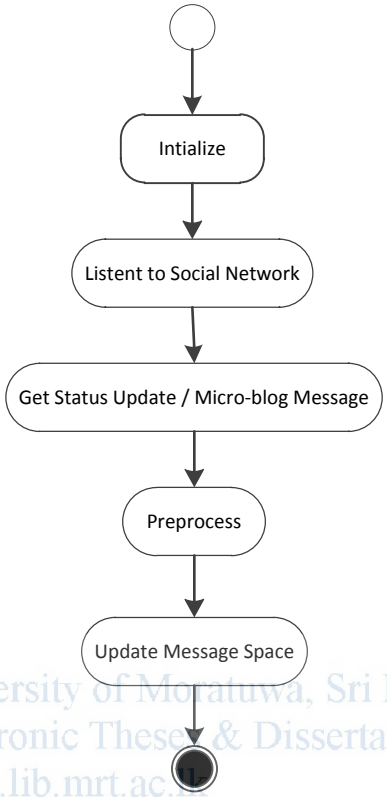

Figure B.2: Activity Diagram of Spy Agent

Figure B.3 shows activities of the crawler agent. Crawler agent is used to crawl web pages to populate the search engine index. Section 6.2.1 in this document discusses more about the crawler agents.

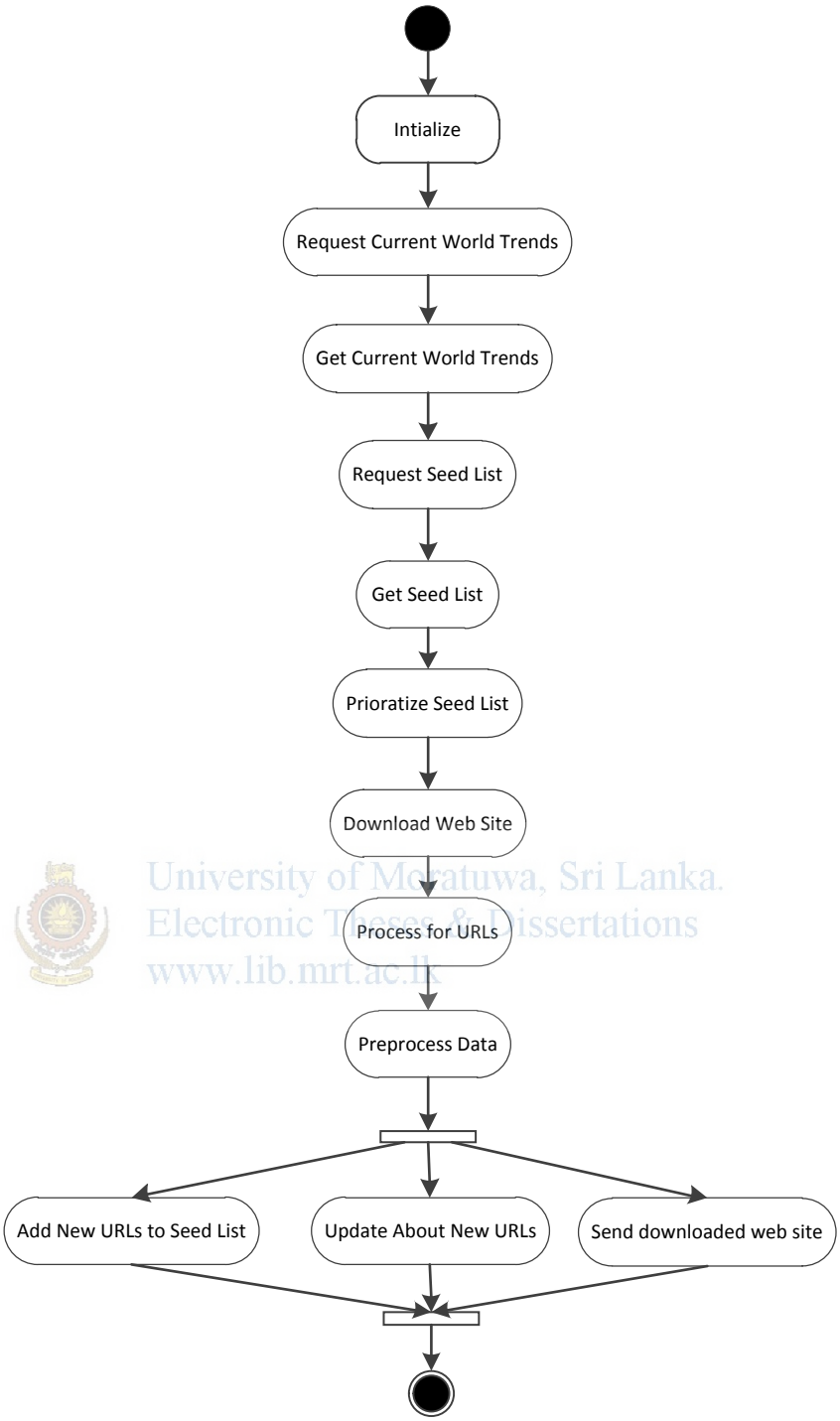

Figure B.3: Activity Diagram of Crawler Agent

Figure B.4 describes the activity flow of world trend agent. World trend agent is responsible for the identification of world trends via status messages and micro-blog messages retrieved by social network spy agents. Section 6.2.3 discusses more about the world trend identification agent used in this project.

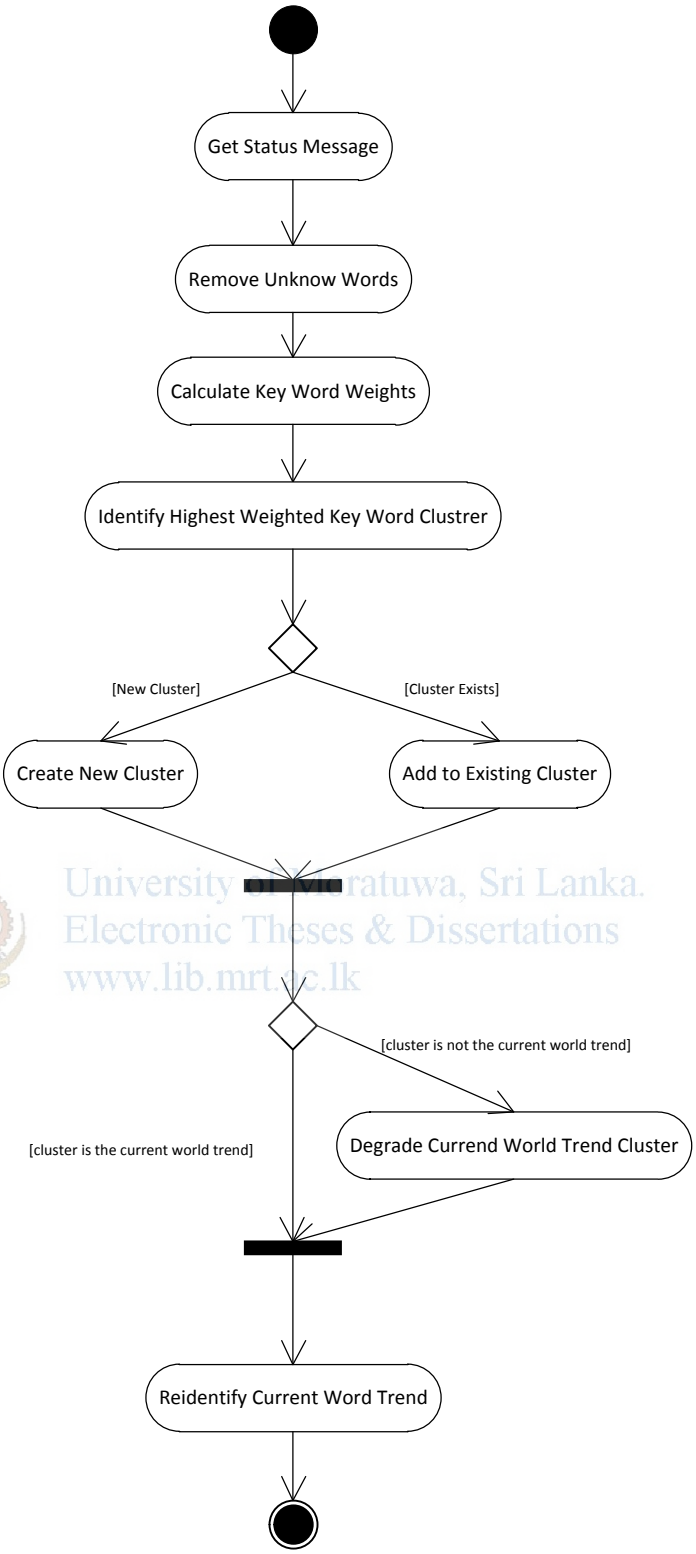

Figure B.4: Activity Diagram of World Trend Identification Agent

Appendix C

Screenshots

This section includes screenshots of the MAS based web crawler system proposed in this project to solve the issue of inefficient information retrieval.

Crawler System Console

Figure C.1 is a screenshot of the crawler system console. Crawler system console logs in system statuses, messages passed between agents, error etc. Section 6.2 discusses about the crawler system proposed in this project.

Figure C.1: Screenshot of Crawler System Console

World Trend Identification Process

World trend identification agent proposed in this project creates tag clouds in JPEG image format to illustrate size, priority of the current topic clusters. The name of the current world trend cluster has the highest font size and font size reduces as the size of the cluster reduces. The figure C.2 illustrates the unfiltered tag cloud of clusters. Figure C.3 shows tag cloud of filtered clusters. Only the clusters with topics available in system ontology will present that screenshot.

Figure C.4 is a combination of several instances of system tag cloud in order to illustrate the evolution of the clusters within the system. Section 6.2.3 discusses more about the world trend identification.

Figure C.2: Unfiltered World Trend Tag Cloud

Figure C.3: Filtered World Trend Tag Cloud

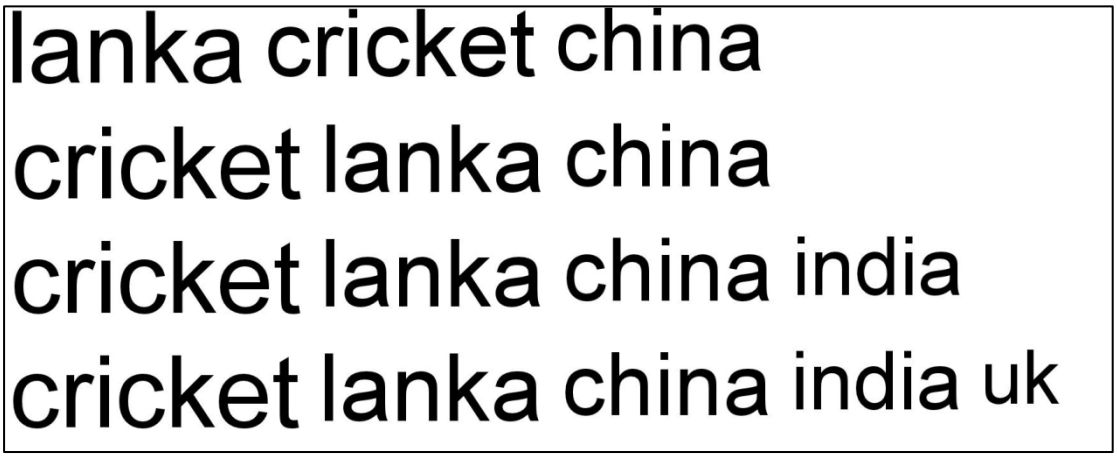

Figure C.4: World Trend Evolution

Search Engine Frontend

Screenshot of the search engine front-end is presented in Figure C.5. Search engine frontend allows users to interact with the proposed search engine. More about search engine frontend is discussed in section 6.3.3.

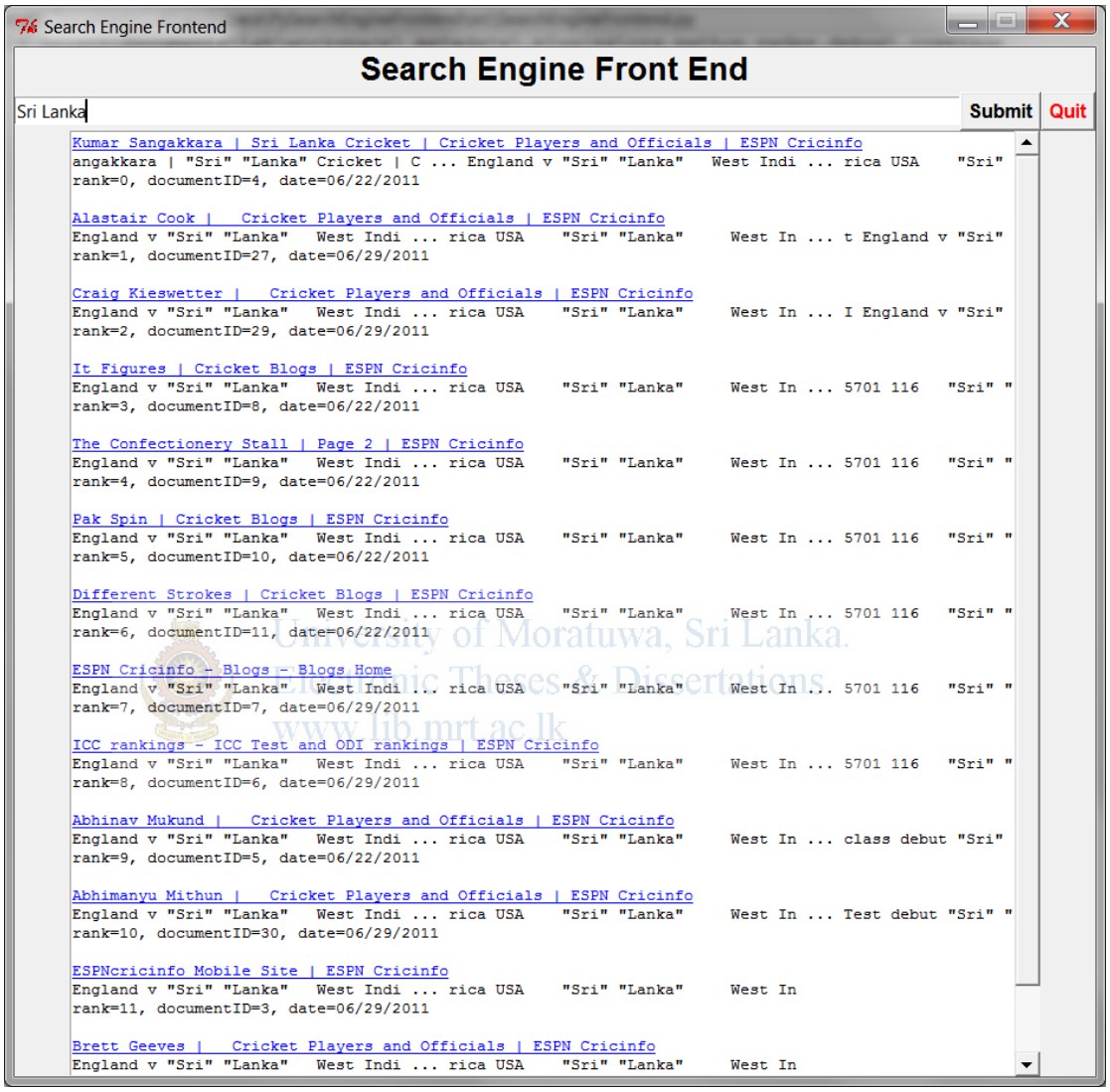

Figure C.5: Screenshot of Search Engine Frontend

Virtual Social Network

Virtual social network is used to manipulate real word social networks for the demonstration purposes in this project. Figure C.6 includes a screenshot of a virtual social network instance. Using virtual social network, user can submit status updates. These status updates will be collected by social network spy agents. More about the virtual social network are discussed under Section 6.4.2.

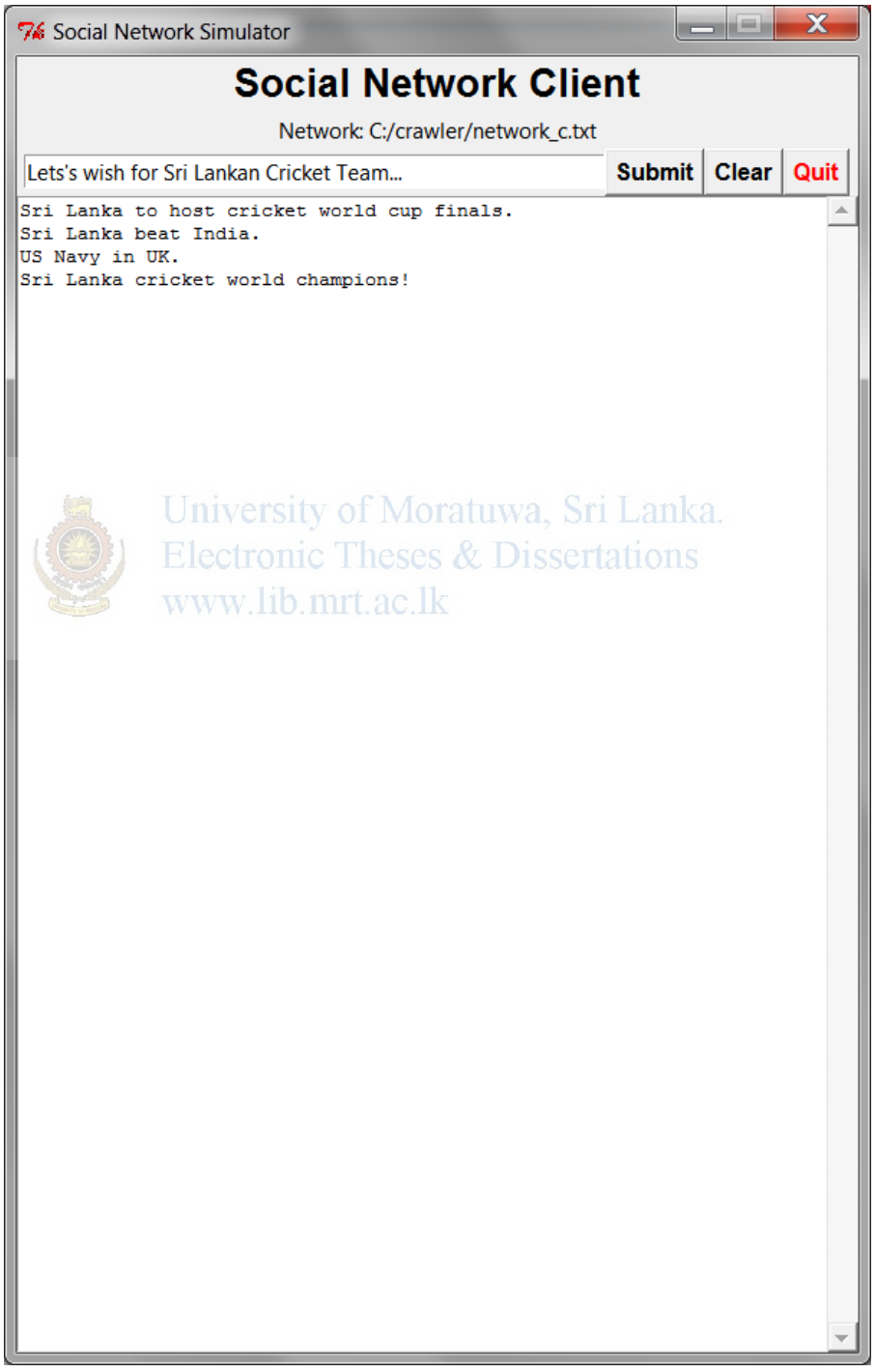

Figure C.6: Screenshot of Virtual Social Network

Appendix D

Sample Seed List Ontology

This is part of seed list ontology used by the proposed MAS based crawler system. The URLs in the seed list have been categorized in to topics. The seed list agent access and make use of this seed list ontology. More about seed list agent are discussed under Section 6.2.4.

cricket>http://cricinfo.com,http://cricket.com,http://cricketindia.com

news>http://cnn.com,http://bbc.co.uk,http://www.news.com

football>http://fifa.com,http://soccer.com,http://football.com

lanka>http://gov.lk,http://www.srilanka.com,http://dailynews.lk

china>http://en.wikipedia.org/wiki/China,http://www.china.org.cn

uk>http://www.direct.gov.uk,www.fco.gov.uk,http://www.england.com

japan>http://en.wikipedia.org/wiki/japan,http://japan-guide.com

automobile>http://en.wikipedia.org/wiki/automobile,http://www.automobile.com

Appendix E

Sample Topic-Words Mapping Ontology

This section includes a sample topic-words mapping ontology. The several words are mapped into major topics. This ontology is used in current world trend identification process. Section 6.5 discusses more about current world trend identification process.

cricket>cricket,wicket,game,match,win,loss,sri,lanka,australia,india,england,new,zeland,south africa,west indies,pakisthan,mahela,jayawardene,kumar,sangakkara,lasthi,malinga,sachith,tendulkar

news>news,world,earth,disaster,bomb,blast,war,terrorism,terrorist,crisis

football>football,soccer,brazil,argentina,england,united,kingdom,australia,maradona,fifa

lanka>sri,lanka,colombo,cricket,mahinda,rajapaksha,matara,galle,kandy,south,asia,buddhism

china>china,begine,beijing

uk>uk,england,united,kindom,queen,prince,william,scottland,london,cricket,football,rugby,soccer

india>india,gandhi,cricket,sachith,tendulkar

australia>australia,cricket,rugby,perth,Sydney

japan>nuclear,toyota,earthquake,tsunami

automobile>japan,europe,toyota,nissan

Appendix F

Test Results for Identification of Updated Web Pages

This section includes test results for identification of updated web pages in the proposed project. Chapter 7 discusses about the evaluation of the proposed project and tests carried out. Table F.1 lists test results for the identification of updated web pages. Topic column shows the topic searched. Delay column shows the time difference between times that web page gets updated and query. “Pass” was recorded when required result was found and “Fail” otherwise.

Topic	**Delay (minutes)**	**Conventional Crawlers**	**Proposed MAS based Crawlers**
Lanka	1	Fail	Pass
Lanka	5	Fail	Pass
Lanka	10	Fail	Pass
Lanka	30	Pass	Pass
Cricket	1	Fail	Pass
Cricket	5	Fail	Pass
Cricket	10	Fail	Pass
Cricket	30	Pass	Pass
India	1	Fail	Fail
India	5	Fail	Pass
India	10	Pass	Pass
India	30	Pass	Pass
Australia	1	Fail	Fail
Australia	5	Fail	Pass
Australia	10	Pass	Pass
Australia	30	Pass	Pass
China	1	Fail	Pass
China	5	Fail	Pass
China	10	Fail	Pass
China	30	Pass	Pass
Football	1	Fail	Fail

Football	5	Fail	Pass
Football	10	Fail	Pass
Football	30	Pass	Pass

Table F.1: Web Page Discovery Test

Table F.2 describes the cumulated success rates of the above recorded results. Chapter 7 includes graphical representations of the Table F.2.

Delay (min)	**Conventional Crawlers – Success Rate (%)**	**MAS based Crawlers – Success Rate (%)**
1	0%	50%
5	0%	100%
10	33.33%	100%
30	100%	100%
Cumulative	33%	87%

Table F.2: Success Rates

Appendix G

Test Results for Efficiency of Web Crawlers

Table G.1 includes test results for the efficiency test of the proposed MAS based web crawler system and the conventional web crawler system. Section 7.2.2 discusses about the efficiency in crawling, evaluation strategies.

	Number of Web Crawls	
Web Page ID	**MAS based Crawler System**	**Conventional Crawler System**
A	1	2
B	1	2
C	1	1
D	1	1
E	1	1
F	1	1
G	1	1
H	2	1
I	1	3
J	1	1
K	1	1
L	1	1
M	2	2
N	1	2
O	1	1
P	1	1

Table G.1: Number of Web Crawls within One Hour