Refreshing Search Engine Index by Swarm Intelligence over Social Networks

Kesara Nanayakkara Rathnayake

09/10010

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

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.

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.

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

List of Figures

List of Tables