Efficient Web Crawler for Forum Sites (ID: 3940)

Tan Ming Sheng 2110022T

# Introduction

Forums have become an important resource on the web due to its increasing richness of user generated information through the contribution by millions of internet users on a daily basis, therefore harvesting forum data can allow analyst to discover useful information to help improve business intelligence. The aim of this project is to develop a user centric crawler to allow them to extract information in a structured way from any websites or forums.

# Progress

Research has been done on related works to identify the current trends in the market followed by the development of a prototype to demonstrate the key ideas of this project. The prototype, codename ‘Focra’ was developed using Django (web framework), Scrapy (crawler) and MongoDB (database), is able to allow multiple clients to start/stop their crawler processes using a basic web interface. Focra is currently under active development and source codes are being uploaded to GitHub at [1].

# Problems

Few problems had occurred during the development of Focra and they are listed below.

1. Lack of up-to-date complex technical resources

Although Scrapy’s website is very well documented, some of the complex technical resources provided by other websites were most using old versions of Scrapy which has already changed drastically.

1. Integrating Scrapy with Django

Scrapy has its own environment settings and it requires at least one or more processes of its own to operate but to integrate it with Django environment and allowing different client process to communicate with their crawler processes was hard.

1. Integrating MongoDB with Django

MongoDB is a document based database and therefore it is different from traditional relational databases. This has caused some minor problems while trying to store data into MongoDB from Django because of the Object-Relational Mapping (ORM) design of ‘models’.

# Next Steps

The next milestone for this project is to enhance Focra with the core functionalities in the order of

1. Allow dynamic selection of data from the websites
2. Provide template for crawling different kinds of website. (Pagination etc.)
3. Allow scheduling of the crawler (when to crawl, how often to crawl)

And if time permits, these following additional functions will also be implemented in no order of

* Controlling the number of crawlers and CPU resource for each client
* Auto detection of website template
* Auto detection of structured data based on client input

# References

[1] <https://github.com/mingsheng36/Focra>