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. In order to efficiently harvest customised data, one would require certain level of technical knowledge to understand the complex APIs and using them is also programming effort intensive. The aim of this project is to develop an interactive user centric crawler system to allow users to visually select and extract usable information (i.e. text, structured data, videos and images) in an intuitive manner from any forums or websites. For the purpose of demonstrating this project, two websites [1] [2] will be used for tests and evaluation.

# Progress

Research has been done on related works to identify the current trends and problems in the market. HTTrack [3] is a website copier and it basically mirrors a whole website by crawling a single or a list of targeted URLs and it also crawls the links inside the targeted URLs which makes it good for offline browsing but it does not allow users to choose specifically what they want inside the targeted URLs, furthermore, it requires installation of the software which means it can only be used locally on the user’s machine and the crawling would stop if the machine is down. A distributed vertical crawler [4] uses a specific set of pre-defined templates for crawling different kinds of websites in a distributed environment however, there are too many websites with different kinds of template and the pre-defined templates may not necessarily be the most efficient way in extracting information of what the users want. It leads to the creation of a prototype, codename ‘*Focra*’ (i.e. Forum crawler) which is a dynamic visual user-defined template crawler to demonstrate the key ideas of this project.

The prototype which uses Django (web framework), Scrapy (crawler) and MongoDB (database) is currently able to allow different users to remotely manage their crawlers by using their browsers and the server will process the crawling for the users. Focra’s development source code has been uploaded to GitHub at [5].

# 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 user 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 Django ‘models’.

# Next Steps

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

1. Allow users to dynamically define their own templates for different websites and forums.
2. Provide different ways of deep crawling for websites and forums. (Pagination etc.)
3. Allow users to schedule their crawlers (when to crawl, how often to crawl)

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

1. Settings to allow users to manage their CPU resource for each crawler
2. Auto detection of website template
3. Auto detection of structured data based on user input
4. User evaluation studies

# References

[1] www.hardwarezone.com.sg - Powered by vBulletin, 2014. [Online] Available from: <http://forums.hardwarezone.com.sg/>

[2] National Center for O\*NET Development, 2014. [Online] Available from: <http://www.onetonline.org/>

[3] HTTrack Website Copier - Free Software Offline Browser (GNU GPL), 2014. [Online] Available from: <http://www.httrack.com/> [Accessed 18 December 2014]

[4] Zhou Bing, Xiao Bo, Lin Zhiqing, Zhang Chuang. A Distributed Vertical Crawler Using Crawling-Period Based Strategy. 2010 2nd International Conference on Future Computer and Communication.

[5] Focra Web Crawler/Scraper, 2014. [Online] Available from: <https://github.com/mingsheng36/Focra>