**Using web scraped data to verify Egyptian consumer price indices**

**Keywords:** Data crawling, Web scraping, Consumer Price Index, Search engine.

# Introduction

The purpose of this paper is to provide an alternative ways of data collection for NSOs, also covers the manipulation and analysis of web scraped data by tracking the utilization of online prices across markets’ websites and cities in near real time.

Recently, In Egypt many companies have been published several websites for e-commerce and one of these is souq.com owned by Amazon, Inc. which made scraping data more available and in general appeared what is called: Web scrapers which are software tools for extracting data from web pages. The growth of online markets over recent years means that many products and associated prices information can be found online and possible to be scrapable.

The consumer price index is one of the official statistics which estimate constructed using the prices of a sample of representative items whose prices are collected periodically; so it’s one of the best examples in this sense: by replacing the scraping of e-commerce websites and websites which publish the currently prices of products to automatically collect prices for some products and services rather than physical visiting to stores to manually collect the prices. This offers a range of great benefits including: Reducing data collection costs, increasing the frequency of collection and products in the basket, and improving our understanding of price behaviour. This paper introduces a developed generic tool that automatically collects online prices, as “Scraped Data”, based on multiple Search Engines to crawler newest prices and e-commerce websites. The developed tool aiming to aid in data collection reduction costs process depend on big data analytics.

Finally, the methodology of this paper is based on machine learning methods that can lead to the crawling of market data on the web, automatic price scraping and evaluation of scanned data.

# Methodology

Our methodology for developing *generic tool* for scraping and analysing big data from the web for official statistics as follows: First we prepare a list of items that we need to have information and prices about them; Get list of web links by passing iteratively item of interest to the search engine; Fetching new links based on the previously gathered one; Add the new links to the seed URL list for online markets; Scrape all the entailed pages.

## The proposed system architecture consists of five main modules. Namely, Data pre-processing based on seed URL list for online markets, Data crawling contains two main functions are searching by Google search engine and fetching the new links, Data scraping and its main function is extraction information from the downloaded HTML pages, Text processing contains two main functions are Tokenize Text and normalize it, and Data structuring, as outline in Fig. 1. These modules are discussed in following subsections [5].

## 

Fig. 1 the proposed Web scraping System architecture.

## Data Pre-processing

## Data Pre-processing phase is the phase of preparing the dataset of e-commerce websites and the other websites which publish or show the prices periodically. Online prices are collected, based on both the website offers and the items' services regarding different regions, from different markets' pages. After then, a customization process is performed for both HTML cookies files and URLs in the markets to excavate the discriminations of prices. Finally, the links associated with categories are stored in a dataset queries for data scraping stage.

## Data Crawling

## Web crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering) so it’s sometimes called a web spider. In our research, we use web search engines especially Google search engine which use web crawling or spidering software to update their web content or indices of others sites' web content by passing new queries for these search engines, which enclose the current date for daily price and web crawlers copy pages for processing by a search engine which indexes the downloaded pages so we can search more efficiently and scrape the prices from these pages.

## So in our research this phase is the process of iteratively finding and fetching web links starting from the URL's seed list. Data crawling must monitor the daily base variety of data and update the URL's seed list regularly.

## Data Scraping

## Using downloaded HTML pages from data crawling phase and e-commerce websites to extract the information. In each page files are arranged in the form of tree nodes annotated as "HTML Tags", where different tags have different both meaning and content. Different HTML pages typically have different tree structures. The downloaded data is processed to fetch these nodes and extract the paragraphs like "Span, div …etc.", prices, which are usually shown with a dollar sign in the form of "<price> </price>" tags, and tables with "table" tag.

## Text Processing

## In CPI, These indices compare prices each month “specific period” with prices in the price-reference month and compares how much it would cost now to do exactly what consumers did in the reference-period with what it cost then; so it’s necessary to unify the specifications of each item to be comparable.

## Text processing is the process of extracting specific text from all of scraped streams. Separating the collected text into a datasets of tokens then get Specific words are observed such as "item name", "units” …etc. in Arabic, same words can be written in different forms, accordingly, a conversion process is mandatory in order to get a unique form for each word. Such unique form is considered as the base form "normalization" for that word, see Table 1.

Table 1 Presents Arabic words, English Meaning, and its base form

|  |  |  |  |
| --- | --- | --- | --- |
| The word forms in Arabic | English Meaning | Base form in Arabic | Arabic pronunciation |
| الأسعار – الاسعار – الإسعار – الأسعـــــــــار – الاســـــــــعار | Prices | اسعار | As'aar |
| المربي – المربى – المــــــربي | Jam | مربي | Morbha |

## Data Structure

The prices which scraped from the websites stored in one structured dataset. And the following information has been extracted and organised in a structured tables:

* **Item Name:** name of the item as displayed on the website.
* **Price:** of the item as displayed on the website.
* **Governorate:** City of the item belongs to, as displayed on the website; if available.
* **Date:** publish date of this price.
* **Type:** information on whether the item is consumer price or producer.
* **Unit:** of the item.
* **Market URL:** in the case, if this item belongs to a particular market.
* **Source:** the URL which scraped from it the price.
* **HTML Source**: the content of the pages.

# Results

The results are summarized in four findings as the following:

## Lack of prices for some products with a limited shelf life such as vegetables, fruits, fresh meat and fresh fish.

## Increasing the number of items which used in the basket especially the electronics products such as the mobiles, computers, home appliances and grocery.

## Decreasing the period between the collection process and show the size of the changes of these items through the weekdays and whole of month.

## By disappearing some items, increasing some else and decreasing the period of collection give great different values compared to the current approach of Egyptian CPI construction.

Thus, the full use of data extracted from the web may lead to a fundamental change in the understanding of price changes in general.

However, it is noted that not all items differ significantly from the current system “manual data collection and static basket items”. Further investigations may be considered to the disappearing some items, increasing some else, decreasing the period of collection or whether the price behavior of these elements differs significantly from the items with large variations. Further the research can be used to verify the accuracy of the manual system until the tool will be completed and can be answer the causes of these results.

# Conclusions

This paper introduces a development of a generic automated tool for scarping and structuring the online data and enhances the process by passing new queries for search engines. This tool offers great benefits in, not only, reducing data collection costs, but also it allows for increasing the number of items without adding any cost; we validated the collected data by that collected by NSO team manually; the tool can allow for increase the frequency of collection, e.g. collect the data daily; discover new items that did not exist in the basket; and aim to understand behaviour of prices. The main disadvantages are: the weights of items cannot be collected from online data; and web data scraping affect the information collected and item representativeness.

# References

1. Hoekstra, R., ten Bosh, O., Harteveld, F. (2012). "Automated data collection from web sources for official statistics: First experiences." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 28(3-4).
2. Robert Griffioen, Jan de Haan and Leon Willenborg, Collecting clothing data from the Internet, Division PIM Department of Methodology (Apr. 2014).
3. Alberto Cavallo, SCRAPED DATA AND STICKY PRICES, NATIONAL BUREAU OF ECONOMIC RESEARCH (Aug. 2015).
4. Robert Breton, Gareth Clews, Liz Metcalfe, Natasha Milliken, Christopher Payne, Joe Winton and Ainslie Woods, Research indices using web scraped data, Office for National Statistics (Sep. 2015)
5. Mina Essam, Mostafa M. G. Mostafa, Safia Abbas, Automating Data Collection for Official Statistics Using Web Scraping, NTTS (Mar. 2017)