German contribution to poster session "Administrative Data, Web Data, 'Big' Data"

Abstract

In the years 2012 and 2013 the FSO conducted a feasibility study on "automation of price collection on internet for multipurpose price statistics".

The main purpose of the project was to implement modern data collection methods that can be applied for both Consumer Price Statistics and Purchasing Power Parities. The feasibility study aimed to explore technical ways, how to gather price data from online retailers.

Quality improvements and efficiency gains can be achieved through an automation of the price collection process via so called web scraping and additional storing software tools. Web scraping is a technique that automatically extracts data from websites at any optional date after once defined the way to this data. Additional software tools for storing allow the saving of the extracted data in external files or databases. In principal an automatic approach can rather lead to significant lower time expenditure for price collections which are conducted periodically in frequent intervals.

The aims of the project were to answer the following questions:

- Is it possible to automate price surveys via internet by imitation of manual collection?
- Is this an efficient survey method?
- What are the advantages or disadvantages?

The findings of the feasibility study were that for a lot of products in CPI and PPP automation of price collection with web-scraping methods is a feasible solution. It's efficient and therefore can help to increase the number of price observations. In consequence this may improve the quality of the respective price indices. Nevertheless, it has to be mentioned that the development of the IT-infrastructure requires profound programming skills. Furthermore, website changes occur irregularly. Thus, support for the adaption of macros and java programs has to be available at any time. Work load for this service cannot be predicted.

The FSO is on the way to implement these collection techniques in the daily production process. But the allocation of stuff resources for support is an essential precondition.

For the existing SPPIs we found fewer products which are suited for price collection with web-scraping techniques. With the enlargement of SPPIs on further industries the situation may change. For some business activities different automation techniques might be more appropriate. E.g. for the services of travel agencies BLS mentioned in the 2014 meeting the use of interfaces to global distribution systems, which we consider to utilize for CPI and future SPPIs as well.

Discussion

After introductory explanations of the feasibility study and the technology, the questions of participants focused primarily on the following topics:

- Does it happen that providers block their websites, if they realize that they are evaluated by web scraping technology? How does the FSO deal with it?

  The FSO informs the providers of the websites about the evaluation for statistical purposes, affirming in the same time that the data is treated confidential and not used for any other purposes than the production of price indices. If the provider of the website is obliged to provide statistical information for price statistics, it might be to the best advantage to allow the evaluation instead of reporting the prices regularly.

- May the price observations be biased by discounts for the first purchase or for frequent purchases?

  In the development of web-scraping methods and technologies in Germany it was intended to imitate the manual data collection of prices in the internet. In case a discount for the first purchase is only promised, the price which is indicated on the website is unbiased. The discount will not come into effect, if the purchase is not executed. Nor can discounts for frequent purchases influence the price collection because the purchases are never executed.

  Other more hidden price dynamics must be condoned. Neither can they be detected in manual data collection.

- Is it assured that quality changes are recognized and how are they treated by the persons in charge?

  In principle, a lot of price determining characteristics of products can be recorded by web-scraping. Nevertheless, at a particular point the technique reaches its limits. The automated procedure is not able to act on modified descriptions as flexible as a human being.