Karola Brunner, Financial and Business Mathematician

Automated price collection via the internet

In recent years, e-commerce has become increasingly important. Based on the outlet type weighting[1] of the consumer price index, e-commerce and mail order business accounted for 5.1 % of the entire basket of goods and services for the base year 2010. For some categories of goods, this share is considerably higher.[2] Prices for the relevant categories of goods are therefore increasingly collected online. Moreover, all major retailers now have their own online shops where they often offer products at the same price as in their local stores. The proportion of products for which price data may be collected online is estimated to be much higher than the figure derived purely from the outlet type weighting.

Automated price collection as part of multipurpose price statistics

For a number of years now, the Statistical Office of the European Union (Eurostat) has adopted “multipurpose consumer price statistics” as a broad heading under which to support projects aimed at modernising price statistics. First, these projects are designed to examine whether, and to what extent, it is possible to collect and use data for the different consumer price statistics (consumer price index, purchasing power parities) on a shared basis, without compromising on data quality. Second, the aim of the projects is to test the suitability of modern methods of data collection for the purposes of price statistics. These methods include the use of mobile devices to collect price data in shops, the use of scanner data and also the use of automated processes to collect data online (so-called web scraping).

The Federal Statistical Office has been conducting a feasibility study on automated price collection via the internet since the start of 2012. Besides analysing the technical possibilities available for collecting price data online via an automated process, it also examines whether this data can be used for the purposes of price statistics. Automated price collection aims to not only reduce the effort and cost involved in this activity, but also to improve the quality thereof. This procedure reduces the number of data transfer errors and, if necessary, sampling may be expanded at very little additional cost. In particular, the time spent on conducting monthly price collections for the consumer price index is expected to fall.

The Statistical Offices of Italy and the Netherlands are also working on projects to automate online price collection. However, the approaches pursued in each project differ from one another. Statistics Netherlands (CBS) has started to conduct initial tests in this field on airline flights.[3] Up until now, the work of CBS has focused on two areas: in addition to conducting tests analysing mass data on individual websites without having a fixed sample, it is also developing a program to simplify manual work carried out in relation to online price collection. For this purpose, price collectors can replicate their sample in the program and are notified of changes in prices and characteristics. The Italian National Institute of Statistics (Istat) is following an

[1] See Sandhop, K.: “Geschäftstypengewichtung im Verbraucherpreisindex” (Outlet type weighting in the consumer price index) in WiSta 3/2012, page 266 et seq.

[2] In particular for clothing and footwear (20.9%), furnishings and household equipment (13.3 %) and for leisure, entertainment and culture (11.1 %).

[3] See Hoekstra, R./ten Bosch, O./Harteveld, F.: “Automated data collection from web sources for official statistics: First experiences”, Statistical Journal of the IAOS 28.3-4 (2012), page 99 et seq.

approach based on a fixed sample which is comparable to the method described in this article. These national statistical institutes share their practical experiences with each other on a regular basis, focusing on technical, functional and organisational issues.

Price collection via the internet

Online data have been used for some time now in consumer price statistics (consumer price index, purchasing power parities). As a general rule, price collectors will access a website manually and enter prices and metadata into files or databases manually. The aim of the project is to examine whether price collection can be automated using web scraping tools. Web scraping is a process whereby websites are automatically retrieved and predefined data are extracted at preset times.

Whether such methods are suitable for the various consumer price statistics depends to a considerable extent on the intervals at which prices are collected and the consistency of the data to be collected. For the consumer price index, prices are collected on a monthly basis. Prices for a specific product are collected up to the point where the product is either no longer important for the market or it disappears from the market altogether. Where this is the case, the product is replaced, i.e. the price collector selects an alternative product and, if necessary, carries out quality adjustments. This product replacement aside, automated procedures generally appear to be a suitable instrument that can be employed for the collection of prices for the consumer price index.

In order to calculate purchasing power parities, the prices for individual categories of goods are collected on a staggered basis every three years. As a result, in contrast to the consumer price index, data for purchasing power parities lack consistency as products need to be comparable in different locations and therefore need to be specified in great detail, although in many cases the specifications are likely to change after three years. For purchasing power parities, automated procedures only seem to be suited to particularly extensive price collections which have to be conducted over a longer period of time (e.g. for flights).

As mentioned above, there are a number of approaches to automated online price collection. The Federal Statistical Office's feasibility study takes the approach of imitating the method used to date of collecting data manually. As far as possible, the steps currently carried out by a price collector are replicated in the program. The samples which have been used to date will be integrated into the automated price collection and, in some cases, extended. Price-determining product characteristics are, if necessary, processed in such a way that they can be examined by an IT program. For example, when collecting hotel prices for the purposes of purchasing power parities, centrally located hotels are to be included. This requirement is met by defining a radius to the city centre. As a result, instead of having to deal with the abstract requirement of having to find “central” hotels, the program will examine the radius. In order to achieve reliable results, practical experience must first be acquired before setting the parameters for such soft criteria.

The feasibility study tested automated price collection for the following categories: flights, hotels, mail order selling (mainly clothing and footwear), mail order pharmacies, hire cars, train travel and city breaks. An automated process collects and verifies online prices and product characteristics. Changes to price-determining characteristics have to be taken into account when calculating the consumer price index and therefore must also be recorded by an automated process. When it comes to purchasing power parities, there is a narrowly defined set of requirements governing the form of the price-determining characteristics, and these must be complied with. Up to now, automated data collections have been tested for prices of both goods and services. Setting up an automated process to collect the price of services is more elaborate than for products given that navigation on service provider's websites is more complex and all steps have to be incorporated into the automation process. The complexity with regard to evaluating price-determining characteristics varies depending on the provider. For instance, as far as collecting hotel prices is concerned, it is relatively easy to evaluate payment and cancellation terms from one provider as these always appear in the same place and their wording is always exactly the same. In the case of other providers, the information is not stored in the same systematic manner, which makes evaluation difficult or even impossible.

Software used

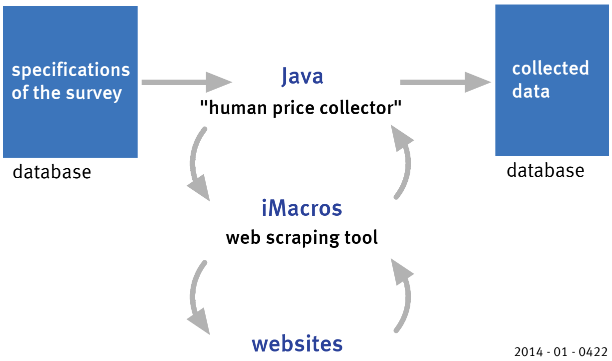
Ease of use was a key criterion in the selection of the software to be used, leading to the choice of the iMacros web scraping tool. The software is similar to a conventional browser. The difference is that it is able to record the individual steps taken while navigating a website. It is able record the filling in of forms and text entered into search fields as well as the navigation on a website. It is also possible to specify areas on the website from which the iMacros tool should extract data. Within the program codes generated by iMacros, which can be subsequently edited, there are always individual elements relating back to the HTML structure of a website which can then be used to aid navigation. Certain sections of the website can be identified using elements and attributes of the HTML document. It is also possible to incorporate texts within the website for the purposes of identification. The reference is selected automatically during recording. However, the elements, attributes and texts used for identification can be specifically defined by editing the code at a later stage. Details of the respective product may be included in the code by using variables.

Depending on the products in question, it is necessary to select data or follow a certain logic when collecting prices. It is not always sufficient for automated price collections to retrieve products using a list and to extract their prices. In some cases, it may be necessary to take product-specific characteristics into account when extracting the price. For example, when collecting the prices of hotels for the purposes of calculating purchasing power parities, there is a stipulation that bookings can be cancelled free of charge and that prices include breakfast. In concrete terms, this means that the price on the website needs to be selected on the basis of these characteristics. The program is designed to verify the data and make decisions. The complex requirements as described above cannot be implemented where iMacros is used alone without any other software.

However, it is possible to control iMacros via a COM interface[4] and to use an appropriate programming language. For the feasibility study on automated online price collection, Java is used to manage the process and retrieve iMacros. Java controls all processes and reads and stores data in a database. Figuratively speaking, Java therefore assumes the role of the price collector. All of the rules that need to be adhered to in price collecting must be predefined. Automated collection lacks the intuition of a human price collector. As a result, rules which have not been explicitly defined cannot be followed. This results in drawbacks whenever the content of a website is changed. A human price collector can spot the changes when visiting the website, but a programme is unable to do this. Such changes may lead to the collection of incorrect data or to gaps in data material. On the other hand, a course of action which follows set rules may also prove advantageous in cases where a human price collector were to select products which he or she personally prefers. A program follows the rules strictly. Where prices are collected manually, details may be overlooked or careless mistakes may occur, especially if the stipulated requirements are complex. In order to access a website, iMacros is started via the COM interface. Commands for navigating the website are also issued via the interface. Where forms or search fields need to be completed, the required data are taken from a database and transferred to iMacros through variables. The extracted data are provided to the Java program by iMacros, also via an interface, and then analysed in the program. The extracted data may also determine the next steps. If, for example, prices are to be collected for a centrally located hotel, a list of hotels is extracted from the website with the aid of iMacros, together with the distance of the hotels from the city centre. The program then rejects all of the hotels which exceed a predefined distance. Depending on the collection being undertaken, several steps may be required until the product and its relevant price is displayed. Once the tool has navigated to the website showing the product data, predefined product characteristics and the price are extracted. Data are adjusted (different formats are harmonised, for example) and, in some cases, initial plausibility checks are carried out. The result is then saved in a database.

The programs are tailored to the individual websites. As a result, for every website on which prices are collected automatically, a certain cost is incurred. However, where prices for a particular category of goods are collected on several websites, the cost incurred does not increase in line with the number of websites. The steps to be taken depend only partly on the individual website. The steps to be taken irrespective of the website can be used to collect other prices whereas steps which are website-specific need to be defined for every single website. In programming, this is done with the aid of inheritance and abstraction. Whenever web scraping tools are used, changes in the structure of a website always incur follow-on costs. Compared to the cost of initial development of the automation process, our experience has shown that these follow-on costs are only minimal.

Figure 1 Scheme of the automated price collection via the internet



Legal situation

The feasibility study also examined whether there are any legal objections to web scraping. In Germany, works are basically protected by copyright. Copyright infringements online often occur in relation to musical works, films or texts. Price collections copy product characteristics and prices from the internet. These are facts and not works so intellectual property is not violated.[5] These facts themselves are not subject to copyright. However, web scraping may constitute a violation of the right to a protected database (Article 87b German Copyright Act[6]). There have been cases where the operators of affected websites have taken legal action against the use of web scraping by commercial users. In one case[7], the flight data and prices of a particular airline were extracted by web scraping tools and offered on another website, the flight tickets were purchased on behalf of their clients and then resold directly. In this instance, the court decided in favour of the plaintiff. The above does not apply to price statistics so this judgement is irrelevant for assessing the lawfulness of the use of web scraping tools to carry out price collections. In three cases, the plaintiffs were online portals or third-party service providers (two airlines and one digital marketplace specialising in the automotive sector). In all three cases, the extracted data for flights and

[4] See *https://en.wikipedia.org/wiki/Component_Object_Model*.

[5] See Sonntag, M.: "Zur Urheberrechtlichen Zulässigkeit von Screen Scraping" (On the admissibility of screen scraping under copyright law) in Schweighofer, E./Kummer, F. (eds.): "Europäische Projektkultur als Beitrag zur Rationalisierung des Rechts" (European Project Culture as a Contribution to the Rationalization of Law), Vienna 2011.

[6] Act on Copyright and Related Rights (Copyright Law) of 9 September 1965 (Federal Law Gazette I, p. 1273), most recently amended by Article 1 of the Act of 1 October 2013 (Federal Law Gazette I, p. 3728).

[7] Higher Regional Court (Oberlandesgericht, OLG) Hamburg, judgement of 28 May 2009, 3 U 191/08 (Regional Court (Landgericht, LG) Hamburg, judgement of 28 August 2008, 315 O 326/08).

cars were used to present the best offers on the extractors' own websites. No purchase or resale took place.[8]

It was examined in the relevant cases whether there was any infringement of Article 87b of the German Copyright Act.

Article 87b Rights of makers of a database

(1) The producer of the database has the exclusive right to reproduce and distribute the database as a whole or a qualitatively or quantitatively substantial part of the database and to make this available to the public. The reproduction, distribution or communication to the public of a qualitatively or quantitatively substantial part of the database shall be equivalent to the repeated and systematic reproduction, distribution or communication to the public of qualitatively or quantitatively insubstantial parts of the database insofar as these actions run contrary to a normal utilisation of the database or unreasonably impair the legitimate interests of the producer of the database.

The court decisions distinguish between two cases, namely whether a substantial or an insubstantial part of the database is reproduced. In the case of automated online price collections, targeted searches are conducted on the basis of a sample. Compared to the range of products offered on the searched websites, the size of the samples are small; consequently, under the judgements handed down by the courts, no substantial part of the contents of the website's database is extracted. Extracting an insubstantial part of the database contents may also be illegal where, in total, a substantial part of the data in the database has been retrieved or where the data is assessed beyond what is considered to be normal use or where it constitutes an unreasonable burden for the operator. None of the above points apply to the use of web scraping tools for price collection. By specifying a sample which remains as consistent as possible over a longer period of time, even repeated searches do not result overall in the extraction of a substantial part of the contents of a database. In the above case, but on other websites too, the operator's general terms and conditions prohibit web scraping of their website. However, general terms and conditions cannot be agreed on unilaterally by the website operator. Only in cases where the user is explicitly asked to accept the general terms and conditions or a contract of use has been entered into in another way must the general terms and conditions be adhered to. In the price collections that have been carried out to date, it has not been necessary to confirm acceptance of the general terms and conditions for the price collections. The circumvention of technical barriers is a problematic aspect (e.g. solving a CAPTCHA[9]) and may be regarded as not being normal website use. However, the tests carried out so far have not faced such barriers.

In general, website operators have the possibility to restrict access to information on their website for price collection at all times. At the moment, only a feasibility study is being carried out. The feasibility study tests technical possibilities and gathers practical experience on using the automated method so as to be able to assess the opportunities and risks it brings. If the intention is to use web scraping for the purpose of price statistics on a long-term basis, clarification would be needed as to whether the permission of the website operators is required or whether a legal basis has to be established. Only then can long-term data access be guaranteed.

Outlook

The feasibility study on automated online price collections will continue to focus on the potential use of the developed methods for the production of consumer price statistics. So far, the requirements in terms of data availability as well as for access to extracted data by the respective members of staff have not been met to a sufficient degree.[10] These aspects are to be improved in future. Changes to websites may result in the data to be extracted being lost. Two measures need to be taken in this regard: changes need to be monitored in a targeted manner and a strategy for handling such technical losses must be drawn up. At present, there are still no long-term experiences on the stability of the method. Furthermore, there is a lack of long-term information regarding the additional cost resulting from changes to websites. More practical experience is needed in order to be able to finally assess the practical suitability and economic viability of web scraping tools for price collection.

[8] Higher Regional Court (Oberlandesgericht, OLG) Frankfurt, judgement of 5 March 2009, 6 U 221/08, (Regional Court (Landgericht, LG) Frankfurt, judgement of 24 August 2008, 2/6 O 478/08), German Federal Court of Justice (Bundesgerichtshof), judgement of 22 June 2011, ZR 159/10 (Higher Regional Court (Oberlandesgericht, OLG) Hamburg, judgement of 16 April 2009, 5 U 101/08, Regional Court (Landgericht, LG) Hamburg judgement of 13 December 2007, 310 O 407/07), Regional Court (Landgericht, LG) Hamburg, judgement of 1 October 2010, 308 O 162/09.

[9] CAPTCHA is the abbreviation for 'Completely Automated Public Turing test to tell Computers and Humans Apart'. CAPTCHA's are used on websites to verify whether the website is being accessed by a human or a machine. This is done in most cases by showing an image containing distorted characters which the user is then required to type into the field on the website provided for this purpose.

[10] For security reasons, price collection is carried out on a computer which cannot be accessed by other workstations.

Extract from the journal Wirtschaft und Statistik

Published by:
Statistisches Bundesamt (Federal Statistical Office),
Wiesbaden

www.destatis.de

Information on this publication

Ellen Roemer
Phone: +49 (0) 6 11 / 75 23 41

You may contact us at
www.destatis.de/kontakt

Statistical Information Service
Phone: +49 (0) 6 11 / 75 24 05
Fax: +49 (0) 6 11 / 75 33 30

Abbriviations

WiSta	=	Wirtschaft und Statistik
JD	=	annual average
D	=	average (for values which cannot be added up)
Vj	=	quarter of a year
Hj	=	half-year
a. n. g.	=	not elsewhere classified
o. a. S.	=	no main economic activity
St	=	piece
Mill.	=	million
Mrd.	=	billion

Explanation of symbols

–	=	no figures or magnitude zero
0	=	less than half of 1 in the last digit occupied, but more than zero
.	=	numerical value unknown or not to be disclosed
. . .	=	data will be available later
X	=	cell blocked for logical reasons
I or —	=	fundamental change within a series affecting comparisons over time
/	=	no data because the numerical value is not sufficiently reliable
( )	=	limited informational value because numerical value is of limited statistical reliability

Figures have in general been roundes without taking account of the totals, so that there may be an apparent slight discrepancy between the sum of the constituent items and the total as shown.