STATISTIK HESSEN

HESSEN

Web scraping from company websites and machine learning for the purposes of gaining new digital data

Working paper

Section PA:
New digital methods,
Scientific cooperation,
Analyses

Table of contents

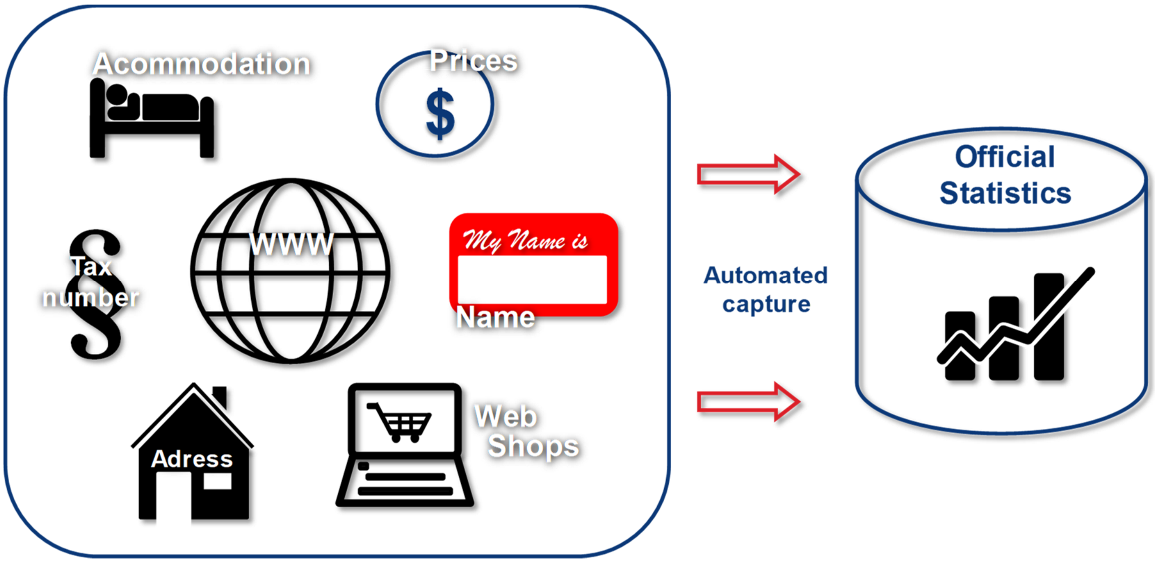
New digital data from the Internet

Company data from the Internet

In 2017, 72% of German companies had a website, 46% were active on social media and 23% had sold goods or services over the Internet (see Statistisches Bundesamt, 2017). Much company data that has, at considerable expense, been gathered, verified and maintained via surveys, interviews or manual research is therefore already available on the Internet and publicly accessible. The necessity of gathering this data via company surveys could be obviated with the use of web scraping.

New digital company data to be found on the Internet is held on commercial websites that generate sales of goods and services. This is especially the case for, for example, electronic order and payment systems such as online shops, commercial online booking portals and service providers. This company information is also available for service providers that have an online presence (e.g. tradesmen, or websites set up for information or marketing purposes – e.g. hair salons). In both groups, which represent the core of the Internet economy, data is generated as a by-product of certain processes and is generally not generated with the express intention of gathering data (see Oostrom, Walker, Slootbeek-Van Laar, Azurduy and Rooijakkers, 2016).

Fig. 1: New digital data for official statistics from the Internet



New digital data can be of various types and covers a wide range of topics. This includes data that, for the purposes of connection with official data, contains important entities or core data such as name, address and tax number, but also important specialist information such as retail item prices, number of hotel rooms available or the availability of online shops for electronic sales. Once found, and provided it is publicly accessible, this new digital data need only be electronically captured and collected via automated process.

Advantages of obtaining Internet data

In various ways, the automated capture of new digital data can lead to the improvement of official statistics:

- **Survey support**: statistical information that currently needs to be collected at considerable expense via surveys or interviews is already available on the Internet and could in the future be gathered regularly and automatically through web scraping. Routine tasks and manual processes would become a thing of the past, which would considerably reduce the cost of data gathering. Examples for this are price-setting as part of the consumer prices index, the determining of bed quotas by hotels or the e commerce activities of Hessian commercial companies.

- **Core data maintenance support:** Because web scraping provides statistically quantifiable information about companies, this information can be used for verification purposes.

- **Speed of data provision and data accuracy**: subject to technical capacity, web scraping could provide data in monthly cycles.

- **Content enhancement**: as well as supporting data gathering processes, web scraping could provide statistically usable information not yet available in official statistics. This includes for example the involvement of companies in e-commerce activities, investment in sustainable technology, security standards on company websites or accessibility for online shops.

- **Less burden on respondents:** publicly available data would not need to be disclosed to official statistics by those under an obligation to do so. This obligation may therefore cease to apply.

In summary, the automated mining of data from the Internet could mean that data and results of the official statistics could be provided in a shorter timeframe, at less cost, with fewer surveys, broken down into more categories and with greater diversity of content. Given our experience so far, we would not expect to completely replace traditional statistics production processes any time soon. But procedures used thus far could be comprehensively enhanced (see Hackl, 2016).

Web scraping – background

Searching for, finding, structuring and storing data

The Internet is often seen as a huge library of digital resources. Underlying these resources some interesting data can be found. The problem is how to find this data reasonably quickly and without great expense. In this context, meta-search engines have become popular. These distribute search queries across several search engines simultaneously and then process the results.

These days many meta-search engines use web scraping. This process involves general processes that request entities from source databases, pass them to meta-search engines and thus enable the websites being searched for to be found. The websites found are then searched for the relevant content, which is subsequently extracted, transformed and linked to other databases (see Salerno and Boulware, 2006).

In this way, web scraping is primarily used to find unstructured information on websites, extract it, structure it into comprehensible formats and thus make it storable for databases, tables or comma-separated text files (see Sirisuriya, 2015).

Web scraping at European and international levels

With the rise of online businesses and online communication, web scraping has already been used to good effect for gathering official statistics nationally. An automated data mining feasibility study carried out for the Dutch office of national statistics (CBS) showed that the collection and processing of data via web scraping is possible, that it can lead to improved learning and efficiency and, especially where large amounts of data are concerned, to faster provision of data and improved data quality, although the costs incurred as a result of process modification and amendments to website infrastructure must be taken into account (see Hoekstra, ten Bosch and Harteveld, 2012).

One of the first common applications for web scraping by national official statistics institutes was the automated collection of consumer prices. The procedure was successfully used to calculate the Argentinian online inflation rate using data from online retailers from 2007-2011. The online inflation rate exceeded the traditionally calculated rate by a factor of three (see Cavallo, 2013). At the European level, the Italian national statistics office (ISTAT) was successfully involved in the automated collection of consumer prices online via web scraping as part of a European project entitled Multipurpose Price Statistics (MPS) (see Polidoro and others, 2015). The German national statistics office (Statistisches Bundesamt) has successfully and increasingly been using web scraping for several years as part of its pricing statistics (see Brunner, 2014 or Schäfer and Bieg, 2016). Web scraping was subsequently extended to other areas of national statistics.

EUROSTAT and national statistics authorities and institutes founded the European Statistical Systems Network (ESSnet) to produce more comparable statistics on a Europe-wide level. Within the network, and following a tender by the European Commission, the ESS-net project Big Data was initiated by 22 national partners. The aim of the project was the integration of big data into European official statistics. It comprises a total of eight work packages involving the mining of new digital data via various methods and channels. Work packages 1 and 2 covered the identification of new digital data through web scraping.

Work package 1, Web scraping job vacancies, was concerned with the automated mining of information on recruitment ads on job portals and company websites. Alongside the Czech Republic, Italy, the UK and Ireland, Germany's national statistics office was involved with its own pilot project as part of the wider European project. This was a feasibility study to capture recruitment ads on recruitment websites (GigaJob.de, Online-Stellenmarkt.net and Jobs.meinestadt.de) aimed at the German employment market (see Zwick and Wiengarten, 2017).

Work package 2, Web scraping enterprise characteristics, was concerned with the automated search, storage, structuring and linking of company websites with datasets from official statistics. The aim was to enrich and improve existing economic and company registers with digital company information. As part of the project, the following company properties were gathered as experimental data for the national registers of companies:

- No. of company websites

- Companies involved in e-commerce
- No. of job offers on company websites
- Companies' social media presence

This involved the national statistics institutes of the following countries: Italy, Bulgaria, the Netherlands, Poland, the UK and Sweden. Germany was not involved in this project. The Italian national statistics institute (ISTAT) was the project lead. ISTAT developed its own Java search routines and, with 78,000 company websites, made by far the biggest contribution to automated information extraction from company websites. The search routines were also successfully deployed by participant countries Poland and Bulgaria.

The feasibility studies that were part of the work packages were carried out from February 2016 to March 2018 and completed with some good results regarding web scraping from company websites and company characteristics. In three pilot projects, the six participating national statistics offices came to the conclusion that, using various web scraping methods, high quality results could be achieved, but that the procedures involved were expensive and there were still many challenges to be overcome.

As part of a new tender by the European Commission for an additional EU-wide research project (ESSnet Big Data II, 2018-2020), another five potential pilot projects were defined. The work package Smart Tourism shall deal with the issue of innovative data sources and methods for tourism statistics. In any event, much of the data relevant to tourism statistics is nowadays available on the Internet in online travel and booking portals and on hotel and tourism websites. If this work package is one of the three projects to receive funding in a European context, web scraping could play a central role as a fundamental method for extracting tourism data from the Internet.

Web scraping within the Hessian statistical office

Since October 2017, the Hessian statistical office has undertaken several web scraping measures and activities. Important orientation aids were provided by the feasibility studies in work package 2 of EESnet's big data project, Web scraping of enterprise characteristics, and especially by the algorithms provided by ISTAT.

The core of the application was the Hessian company register containing important core data from around 300,000 Hessen-based companies. The aim was to find the websites of companies and connect to and use their publicly available data.

The algorithms for finding, reading, structuring and linking the data available on company websites were applied on a sample of Hessian companies from the database of official statistics, and with good results. As can be seen below, the first stage saved successful connections made to the websites of various Hessian companies. There then followed an evaluation of part of the tags readable on the linked-to websites using text mining methods. Finally the existence of an online shop for Hessian companies was established using methods of predictive modelling with the assistance of training data and an input/output function.

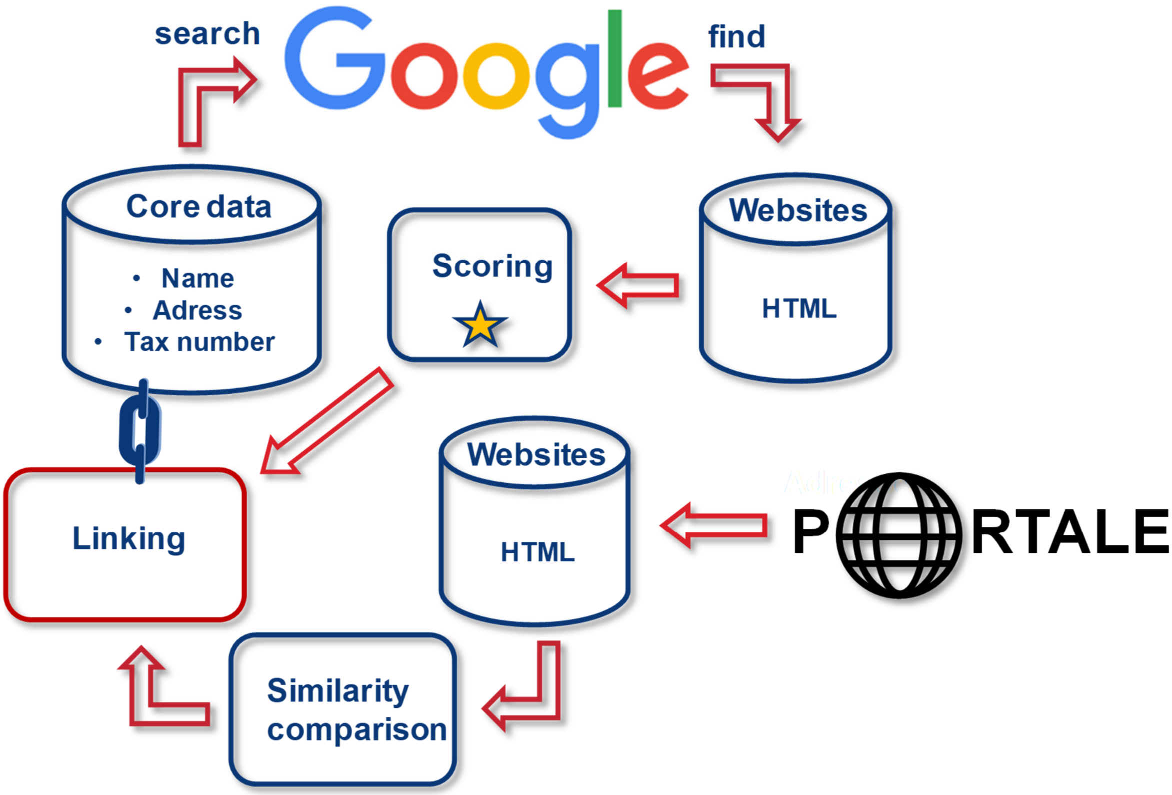
Procedure – search, find and link

The search for quantifiable content on the Internet using information extraction is often carried out by webcrawlers. This involves all searchable web pages being methodically searched and saved. Web scraping involves searching for specific, pre-defined information and the structuring and extraction of quantifiable content. In this case, existing companies based in Hessen are searched.

Web scraping uses identifying characteristics or pre-information in searching for specialist content and links. Publicly available data is frequently unstructured and needs to be transformed into a form appropriate for reading (see Vargiu and Urru, 2013).

As in this project quantifiable information was to be mined from company websites, register-based core data saved on official statistical databases and publicly available data in commercial online booking portals was treated as pre-information. Core data that companies can identify contains addresses, company descriptions, legal structures and information on subsidiaries and branches. This information can be used as datasets for Internet searches. At the same time, data tables mined from databases are suitable for being enriched with new digital information (see Fig. 2).

Fig. 2: Linking websites using Google and portal searches



Scraping company websites using meta-search engines

Fully automated web scraping is suitable for identifying, storing, structuring and using data and quantifiable content from official statistical databases with available company information and for linking it to official statistical data.

In view of ISTAT's comprehensive and successful activities in the automated extraction of information from company websites, the HSL (State Statistical Office of Hesse) has closely aligned itself with the procedure that ISTAT has provided (see Barcaroli, Scannapieco and Donato, 2016).

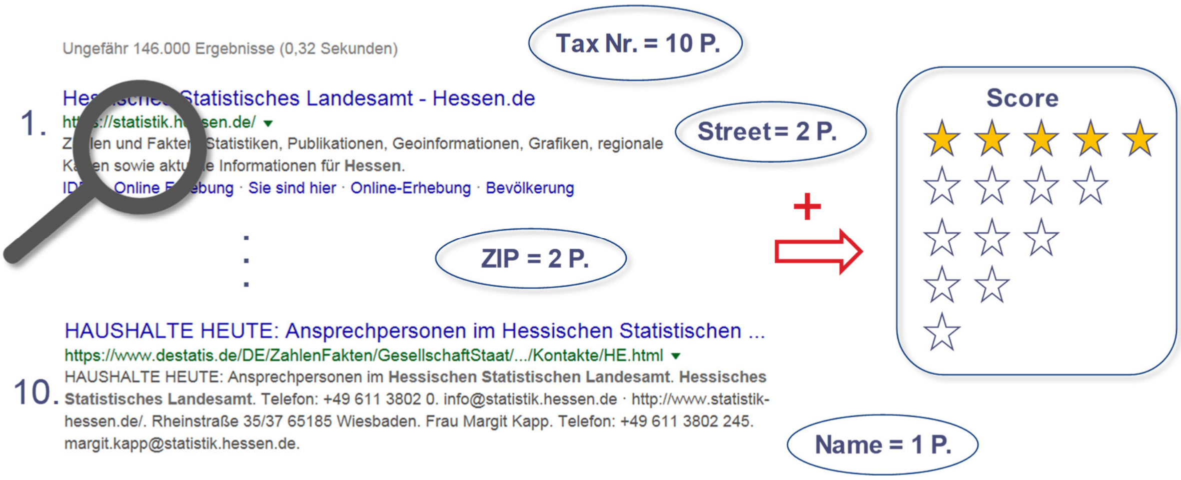
The datasets to be enriched, that contain identification features/entities such as name, address or tax number, are initially saved in a database. From there, an algorithm written in the programming language Java accesses the individual datasets in order to use them for additional searches using a URL crawler. In order to search the webpages for the identification features, a meta-search engine such as Google must be used.

For Google searches for company websites, the identification features from official statistical sources are used as datasets. Following input of a company's features, the Google search results page shows a specific quantity of Internet references found.

Subsequently, the source texts/webpages of the ten most highly placed company websites (main web pages) as well as the identifiable secondary webpages contained therein (company information, About Us, contacts us etc.) and the Google knowledge panel (Google right-hand page) are saved and searched in accordance with suitable core data. In this way up to 44 web pages were searched per input/company.

Depending on quantity, type and expression of the identification features found, the main and secondary pages saved are now evaluated using a scoring system. The marks awarded are summed across the secondary pages and provide a weighted accuracy score. Then the websites are sorted according to the amount of the score and their Google ranking. Then the main company page with the best ranking/score is assigned to the corresponding core data. This two-stage sorting ensures one-to-one sorting. In this way, the number of found and allocatable webpages per company is reduced to one.

Fig. 3: Evaluation system



Ungefähr 146.000 Ergebnisse (0,32 Sekunden)

1. Hessisches Statistisches Landesamt - Hessen.de
https://statistik.hessen.de/
Zahlen und Fakten, Statistiken, Publikationen, Geoinformationen, Grafiken, regionale Karten sowie aktuelle Informationen für Hessen.
IDEV Online Erhebung · Sie sind hier · Online-Erhebung · Bevölkerung

10. HAUSHALTE HEUTE: Ansprechpersonen im Hessischen Statistischen ...
https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/.../Kontakte/HE.html
HAUSHALTE HEUTE: Ansprechpersonen im **Hessischen Statistischen Landesamt**. **Hessisches Statistisches Landesamt**. Telefon: +49 611 3802 0. info@statistik.hessen.de · http://www.statistik-hessen.de/. Rheinstraße 35/37 65185 Wiesbaden. Frau Margit Kapp. Telefon: +49 611 3802 245. margit.kapp@statistik.hessen.de.

In this way, it is possible to link all the publicly available information on the allocated webpages, insofar as it is quantifiable. It could have been the case that none of the eleven resulting webpages was allocatable to the search engine. In this case, it would not affect the points sequence but rather the overall number of points. That the allocated webpage was initially "right" was at this point automatically checked by ISTAT using a probability-based R-programmed machine learning process. For HSL, a rule-based approach proved to be sufficient. This contains a minimum point score of 5 to be reached as the first initial truth test.

Scraping of commercial online portals

Many small companies often do not have their own website but are represented on a commercial online portal. For these searches, the automated finding, saving, structuring and linking of data works differently from web scraping by means of meta-search engines. The process developed by ISTAT tries to download the company information page in order to look for identifying features. With commercial portals, this method will find only the portal operator, not the company concerned. For this reason, as part of a field test, HSL developed and programmed its own algorithm for information extraction from a commercial online booking portal.

Online portals can be seen as specialist (e.g. recruitment, accommodation, property etc.), commercially used indexes/compilations of various companies in their function as providers of a specific product or as part of a specific sector in various regions. Generally these companies want to be found on this portal by potential customers. Small companies and freelancers without their own website are especially registered on such portals.

Knowledge of the sector or industry in which the company operates therefore makes it possible to automatically scan portals for companies prior to, during or after the web scraping linking process. By means of URL crawling, the portal will be searched not for the company concerned but for all companies of the sector concerned in the city concerned. Once found, the URLs of the relevant portal sub-pages can be linked to the official statistical datasets.

Done in addition to web scraping, portal scraping has the following advantages:

- **Capture beyond relevance thresholds:** with portal scraping, companies can be captured that, due to specific sales figures or employee numbers, are not contained in the core data of official statistics. This could relate to freelancers or small companies not registered for sales tax.
- **Efficiency:** the number of webpages to be searched is several times smaller than it is with web scraping using a meta-search engine, as portals contain only websites of the sector concerned. Portal scraping requires less storage space and has proven to be many times faster than the web scraping of company websites.
- **Less complex:** The number of work steps involved in portal scraping is smaller, which makes it simpler than the web scraping of company websites. Company features and core data are always structured the same way on the various sub-pages on the portal. This means that the search algorithms are easier to program.
- **Accuracy:** Thanks to its relevance to the desired sector and region, the reading and allocation of digital information from online portals has a high number of relevant hits. It is therefore not necessary to assess the accuracy of hits. For linking with core data from official statistical sources, the hits need only be compared for similarity.
- **Data frugality:** For portal scraping, the automatic input of core data into a search engine is not required until the linking stage. Portal data can however be gathered and processed/used prior to linking. With web scraping, searches for corporate information without core data can only be undertaken with difficulty.

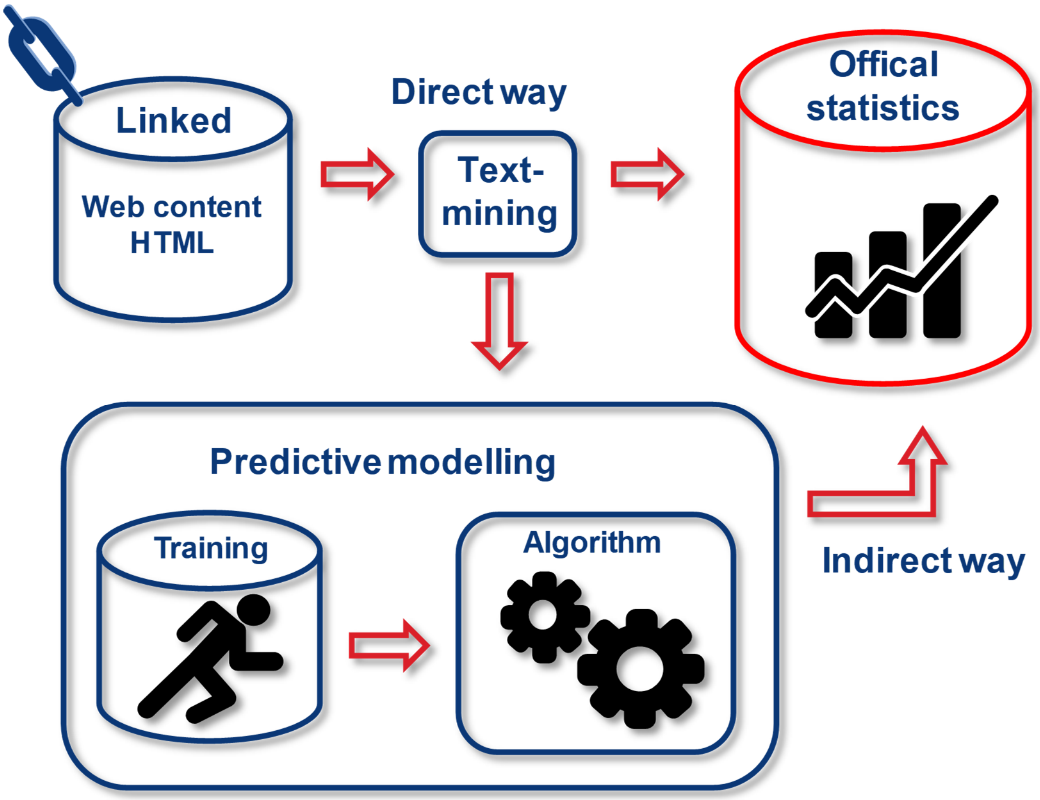
In order for portal scraping to be carried out, suitable portals must be researched manually for relevance prior to automated data gathering. Depending on the subject matter, the number of suitable online portals is however limited and more or less searchable. The efforts and expense involved in the manual research are therefore seen as reasonable.

For the linking of content found, a similarity comparison must be carried out. For this one can use easily implementable R- or Java-programmable metric word distance measures, such as the Levenshtein Distance, the Monge-Elkan Distance or the Jaro-Winkler Distance. These are based on the simple comparison of characters and letters of different words (see Cohen, Ravikumar and Fienberg, 2003).

Data enrichment with linked content

As shown in Fig. 4, data obtained via web scraping can be linked in various ways with official statistics. This involves direct transfer whereby the data obtained is imported into the database of official statistics in a structured way as indicators, keyword counters or readable feature indicators.

Fig. 4: Potential for enrichment of datasets by web content



Where desired company characteristics cannot be assessed via simple reading, there is still the option of data enrichment using machine learning processes (indirect path).

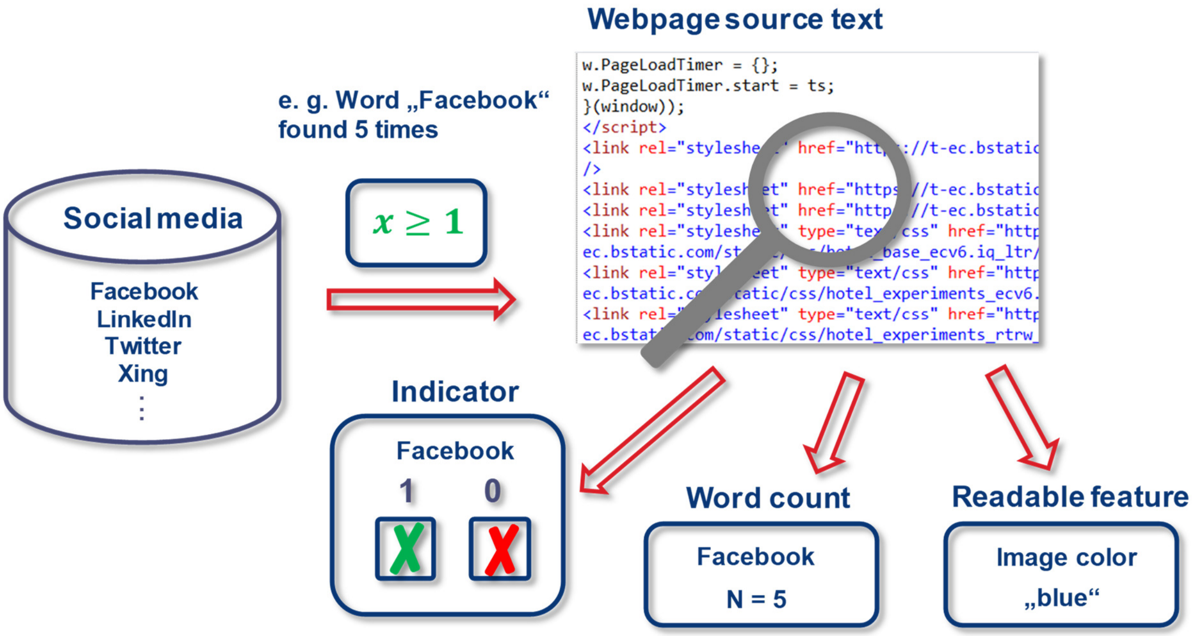
Direct data enrichment

If the websites concerned (company websites and company-related secondary portals) have been linked to the statistical units (from database) of official statistics, these can be searched for specialist content.

First of all the relevant source texts are downloaded. The data contained in the source text is not yet quantifiable information but rather is available solely in unstructured or variably structured form. The options for turning the unstructured data available on the company websites into quantifiable, structured statistical information are as follows (see also Fig. 5):

- **Keyword-dependent generation of indicators:** When the desired keyword appears, the word indicators are assigned the value of 1. If the word in question is not found on the website, the word indicator is assigned the value 0.
- **Keyword counter:** this states the frequency of the searched-for keyword.
- **Readable features:** where feature indicators on the website are readable (e.g. number of double rooms on the website of a hotel or the logo colour of Facebook) then the feature can be expressed as a direct variable.

Fig. 5: Structuring unstructured source text data



```
w.PageLoadTimer = {};
w.PageLoadTimer.start = ts;
}(window));
</script>
<link rel="styleshe[illegible] href="htt[illegible]://t-ec.bstatic
/>
<link rel="stylesh[illegible]et" href="https[illegible]//t-ec.bstatic
<link rel="stylesh[illegible]t" href="http[illegible]//t-ec.bstatic
<link rel="styleshe[illegible]" type="tex[illegible]css" href="http
ec.bstatic.com/st[illegible]_base_ecv6.iq_ltr/
<link rel="sty[illegible]eet" type="text/css" href="http
ec.bstatic.c[illegible]tatic/css/hotel_experiments_ecv6.
<link rel="[illegible]lesheet" type="text/css" href="http
ec.bstat[illegible]com/static/css/hotel_experiments_rtrw_
```

The precondition for direct data capture in downloaded source texts using the three methods described is the context-dependent appearance of keywords. The keyword "single room" can appear on hotel websites several times in multiple contexts, such as in the comments left by a guest. For the automated capture of the number of single rooms in the hotel concerned, it is important which characters, words and templates appear before and after the keyword.

Here the use of regular expressions as a method of text mining is unavoidable for structuring in, for example, R (see Munzert, Rubba, Meissner und Nyuis, 2015).

To transfer data in the way described into the datasets of official statistics is seen as a direct data capture method via the Internet following linking.

Indirect data capture – automated assessment of company properties

Specific company properties such as e-commerce activities are indicated by the use of online shops for the electronic sale of goods and services. The presence of an online shop cannot always be read directly from a website, even if the website contains one. Statistical methods for determining latent properties can help to make such properties visible and to classify the websites as binary. The precondition for such classifications is that websites that contain an online shop must differ from, for example, pure online presences for marketing purposes.

Specific features that are present by way of the structuring shown can be related to the searched for company property. For the property of "online shop", this can apply for the presence of a payment system, a purchase system, a social media link or trade in specific goods. Insofar as one or more such features are present on a website, the site can be classified as containing an online shop. This method can be described as "company feature allocation in accordance with deterministic decision rules".

The disadvantage of this method is that the connection between the features and the searched for property must be known and the decision rule within the search cannot be changed or adapted. It must be decided in advance how many features determine specific properties of an online shop and in what way. It is then however possible that the features described also appear on websites that cannot be classified as containing online shops. In this case, the use of a deterministic decision rule would be subject to error.

Predictive modelling

Predictive modelling is a popular procedure of machine learning which enables the estimated causality between the captured features and the probability of the appearance of the desired company information to be viewed as a basis for classification. Common areas of application are the insurance industry and business intelligence, where predictive modelling algorithms are used for segmenting customers, predicting sales, analysing markets and assessing risk (see Frees, Derrig and Meyers, 2014).

This process is also frequently used for online marketing, spam identification, fraud prevention and customer relationship management (identification and segmentation of potential customers according to likelihood to purchase). Using historical data, it can help determine which product types might interest users or on what fields, buttons and links they are likely to click (see Tuzhilin, Gorgoglione and Palmisano, 2008).

Function learning

Predictive modelling procedures are based on the probability of the appearance of the property whose causal relationship to the captured features is represented by the unknown function f. The interesting property of this function is nominally scaled and therefore has a Boolean output value and is described as a classifier. The Boolean value comprises a positive case when the interesting property appears (e.g. online shop) and a negative case when it does not appear. A hypothetical input-output function h is now defined. The form of this function is arbitrary and here follows a logistical distribution.

$$h = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

The parameters β stand for the causal relationship between the capture features in X and the likelihood of appearance. The output of the hypothetical function provides the estimated probability of the appearance of the interesting property. Thus the methodical approach used here is based on logical regression within the learning algorithm (see Long, 1997).

There are various methods for predictive identification but logistic regression is nonetheless a very popular and easily comprehensible method that has especially proven its value in the area of pattern recognition in medical IT. It is based on probabilities and is easy to apply (see Dreiseitl and Ohno-Machado, 2002). In identifying online shops, ISTAT used the machine learning algorithms Neuronal Networks, Logistic Regression and Random Forest and checked the results by measuring for precision, sensitivity and accuracy. It was thereby apparent that, as an algorithm of predictive modelling, Logistic Regression created no greater error rate than the significantly more arduous and complicated alternative processes.

Training regime

The aim of function learning is now to achieve ideally identical results using h as are achieved with function f. This is achieved by applying the hypothetical function to historical electronically gathered company data with already known company properties that are distributed among a training dataset and a test dataset. Using the procedure of declining gradients, the causality parameters are repetitively specified/learned with the training data via the minimisation of a convex empirical error function derived from the hypothetical function with the training data and checked using the test data. This type of function learning is called gradient-based learning. The use of training data for fully supervised function learning is frequently done in batch mode. This means that all data sets of the training data that have been manually researched beforehand are used in an optimisation process. The more cases are classified as correct, the better the function has been learned.

With the approach outlined here, the repetitive online training regime is applied as part-supervised function learning. This means that the training and test database initially comes solely from manual research but is not static. The database is rather strengthened via the automated involvement of new datasets, depending on the test results of the applied predictive procedure. The number and type of the new training datasets follows a function (see Bottou and Le Cun, 2004).

Following the learning process, datasets of linking data with very high or very low predicted probability for the appearance of the company property are automatically added to the training data as part of an HSL process and the predictive process is run again. In this way there arises a repetitive learning process via which the causality parameters can be adapted procedure by procedure.

Prognosis

With the learned causality parameters from the training and test processes and the hypothetical function, the probabilities of the interesting properties appearing can now be specified for unknown data following electronic extraction of the capture criteria. If the probabilities exceed a predefined threshold value, the interesting property is automatically determined for the website (see Nilsson, 1998).

Application to official data

Technical implementation

From October 2017 to May 2018, an IT infrastructure suitable for web scraping was built at HSL.

This involved the installation, deployment and development of a Java program provided by ISTAT for the search and extraction of websites via meta-search engines and for the saving and assessment of the source texts therein.

For this purpose, a database with the corresponding official core data was built on a virtual server at HSL. For the linking and processing of the company websites found and for enrichment with specialist data, various R programs were developed.

Core data from the official company inventory

For the creation of official statistics, companies are usually selected from the inventory of statistical company registers and asked for appropriate operations-related content relating to specific specialist areas.

The statistical company register is a regularly updated database with companies and business from practically all sectors with taxable sales deriving from supplies and services and/or employees. The sources of the company register include administrative data from administrative areas such as the federal employment office or the financial authorities as well as information from individual departmental statistics such as those captured by producing industries, trade or the services sector (see Statistisches Bundesamt, 2015).

The statistical company register is the basis for almost every official economic statistic and contains important core data such as name, address or tax number from some 340,000 Hessen-based businesses.

As current technical capacities allow the capture and linking of 100 corporate websites a day, the automated capture of the entire Hessian statistical corporate register is not feasible within a reasonable timeframe. But as a current random sample of the Hessian corporate register, the 1,658 units surveyed as part of the information and communications technology survey (IKT 2017) were nonetheless sufficient, as these companies were asked i.a. about whether they have their own website. The company features were loaded onto the database on the webserver and the web scraping procedure was applied to the datasets.

Outcomes of linking

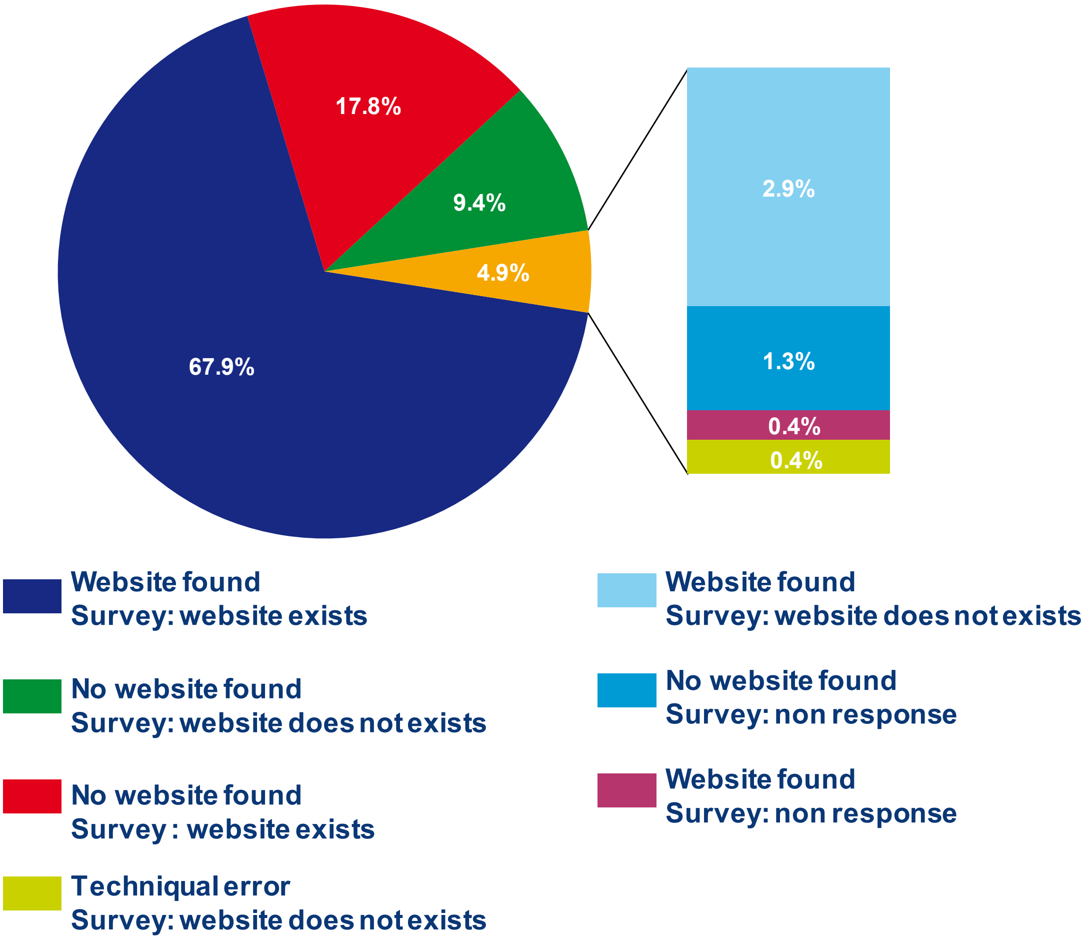
Fig. 6 shows the outcomes of information extraction from web scraping using the inventory of the 1,658 companies from the company register that were surveyed for IKT 2017 with regard to their information and communications technology and the maintenance of their own web presence. The subject of the illustration is whether a web presence has been assigned and whether the company concerned has specified their own website in the IKT 2017 survey.

For 77.3% of the surveyed companies, the results of the survey accorded with the results of the web scraping. 67.9% had a web presence and 9.4% did not. Of 1,420 companies who stated they operated a website an allocation was achieved for 79.2%.

For 21.1% of the companies the web scraping allocations differed from the survey results. 18.2% were unrecognised web presences (17.8% not recognised and 0.4% technical failures) and 2.9% achieved allocations, although in the survey it was stated they there was no web presence. 1.7% of the companies gave no information. Out of 211 companies who stated they had no website, an allocation was nonetheless achieved for 22.7% of them via web scraping.

In 48 cases, companies stated they had no web presence, although such a presence was established on checking and correctly allocated via web scraping. As an explanation for the discrepancy between the survey results and the web scraping, the differing capture times cannot be excluded.

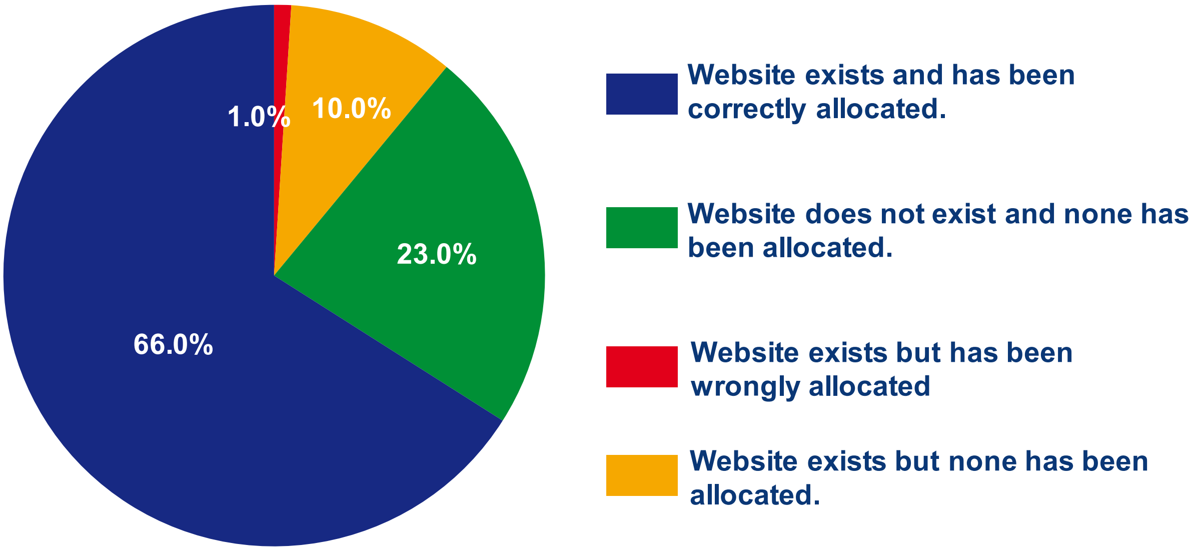
Fig. 6: Outcomes of allocation of extracted websites to 1,658 Hessian companies from the official database



Overall 1,186 web presences were therefore automatically assigned for 71.5% of the companies. Regardless of whether companies stated that they did or did not operate a website, it had to be evaluated whether the allocations were correct. The allocation was therefore checked via manual research for a random sample probe of 100 companies. Fig. 7 illustrates the results of this check.

Correct allocations were established in 89.0% of cases examined (accuracy rate). 66.0% were recognised as without online presences and 23.0% were companies with their own website. The researched web presences of companies with their own website were correctly recognised in 85.7% of cases (sensitivity rate). The automated positive allocations of company websites thus turned out to be correct in 98.5% of cases (precision rate).

Fig. 7: Results of checking of 100 allocations of extracted company websites



In 11.0% of cases the process led to false allocations (error classification rate). In 1.0% of cases, the presence of a corporate website was recognised but the wrong online presence was assumed and allocated. In 10.0% of cases, no web presence could be found.

The various inspection results of the binary classification make it necessary, when interpreting the results, to be able calculate a value for the assessment of the allocation values as a relative correctness intensity. This is the F value, corrected as the harmonic medium between sensitivity rate and precision rate.

The F rate is 0 at the lowest correctness and 1 at the highest correctness. When applied at HSL, it was 0.917, which suggests a very high correctness of the classification carried out.

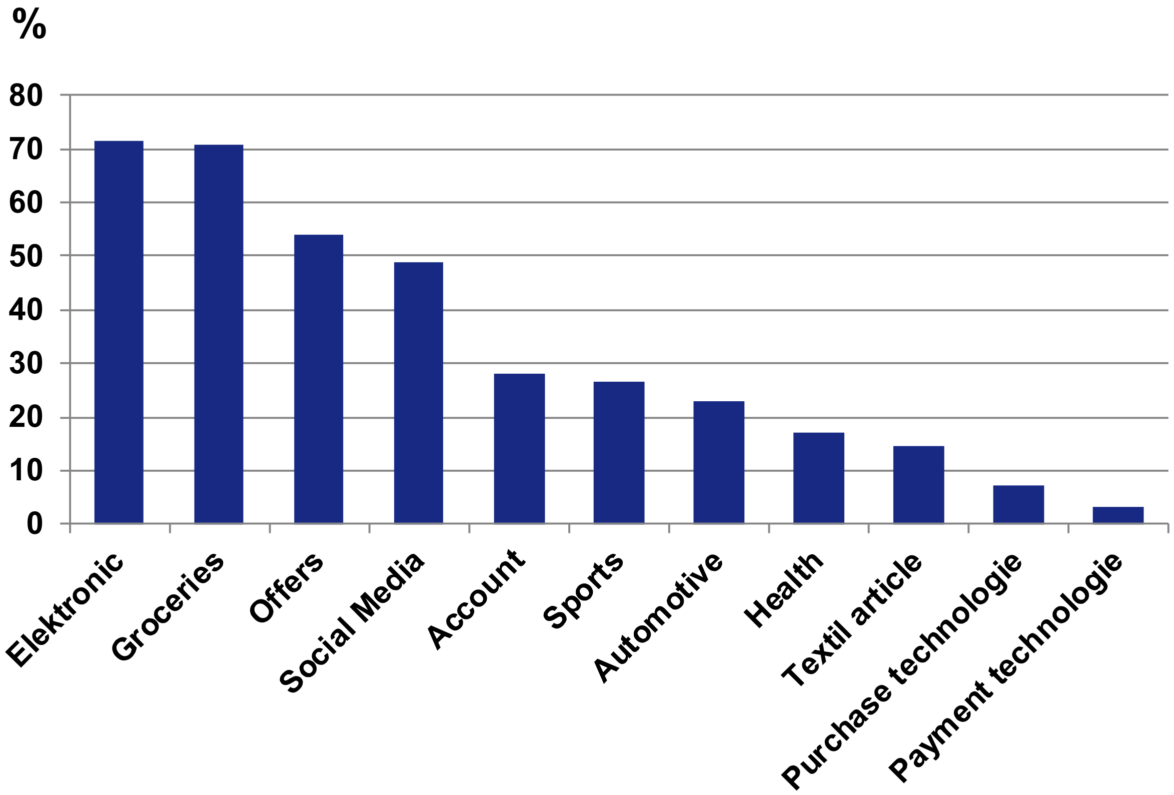
Of the 1,186 identified and linked websites of Hessian companies in IKT 2017, 1,111 could be downloaded and their source text analysed. This involved the compilation of keywords related to "trade" into various groups and the checking of their occurrence in the downloaded source texts. The presence of words from respective word groups was established within various indicator variables.

The results are shown in Fig. 8 and show that most keywords were found for the field of “electronics” (71.5%) and “food” (71.0%) followed by words related to “special offer” (54.2%). In around half (49.1%) of all analysed websites, references to social media were found.

Further keyword group occurrences with frequency of over 10% were in the areas of “account technology” (28.0%), “sports articles” (27.0%), “automotive” (23.0%), “health” (17.1%) and “textiles” (14.7%).

The keyword groups of payment technology (Visa, credit card, PayPal, EC card etc.) and purchasing technology (basket, shopping cart and similar), which are strongly connected to the feature of having an online shop, were found in less than 10% of all allocated company websites.

Fig. 8: Appearance of selected keyword groups on n = 1,111 readable websites of Hessian companies in IKT 2017



Automated determination of the latent company characteristic of e-commerce

There are company characteristics, such as involvement in e-commerce activities, that are of significant interest for official statistics. This characteristic is fulfilled by the presence of an online shop. Often the first view of a website with an active online shop on which orders can be placed and paid for makes it clear that the characteristic “online shop” cannot be read immediately from the website. The operation of an online shop for electronic sales is therefore not directly and unambiguously readable electronically from the source texts of the allocated company websites. There are nonetheless characteristics, such as the presence of online payment operations or a shopping cart, that could reasonably be assumed to relate to the presence of an online shop.

To carry out automated predictive investigation for the latent characteristic "online shop", appropriate predictors must first be found among the automatically captured characteristics in order to support the connection hypothesis referred to. The availability of historical data in which the characteristic "online shop" was already recognised was also necessary. For this purpose a random sample of initially 146 companies was drawn from the data captured. Then, using manual research, a binary classifier was generated that contained a positive case for the identification of a webshop and a negative case if no webshop could be read from the website concerned.

An association analysis with historical data resulted in significant connections between the presence of an online shop on the linked company websites and the following characteristics:

- Presence of purchasing technology (e.g. shopping cart, basket, login)
- Presence of payment technology (e.g. Visa card, PayPal, bank transfer)

With the predictors, a predictive probability-based model, as described in the previous points, was now selected as the machine learning procedure for the automated detection of online shops. For this purpose, the historical data was randomly divided 50-50 into training samples and random samples.

Using the predictive model, the causalities were now assessed between the predictors and the classifier for the property "webshop" in the training random sample in the first learning process and tested using the test random sample. Then, using the learned causalities from the database of the linked-to company websites, 15 datasets with an estimated probability for the property "online shop" of at least 70% were selected from the learned causalities and added to the historical data random sample. Then the entire learning process was repeated for a new learning process with a new database comprising training and test random samples from 161 observations.[1]

With the estimated causalities from two learning processes, it was possible to carry out a classification for all allocated 1,111 websites from companies in the IKT 2017 survey in accordance with the property "online shop". An online shop was predicted for around 5% of the company websites.

For the 161 tested allocations of online shops following two learning processes, the results were again inspected. Figs. 9 and 10 show the results of this test.

For learning process 1, correct predictions were achieved in 87.7% of inspected cases (accuracy rate). 5.5% were recognised online shops and 82.2% were company websites without an online shop. The online shops researched were recognised correctly in only 30.8% of cases (sensitivity rate). The proportion of correct allocations of online shops to all positive classifications was nonetheless 100.0% (precision rate).

In 12.3% of cases, the procedure led to false predictions and failed to recognise existing online shops (error classification rate).

The F value here was 0.471, which suggests significantly less accuracy in the identification of webshops carried out than was achieved for the allocation of company websites.

For learning process 2, in 91.4% (+3.7%) of cases reviewed, correct predictions were made (accuracy rate). 13.6% were recognised online shops and 77.8% were company websites without an online shop.

The researched online shops were now recognised correctly in 61.1% of cases (sensitivity rate). The second learning process thereby increased the sensitivity rate and thus the recognition of websites by 30.3%. The proportion of correct allocation of online shops to all classifications was nonetheless 100.0% (precision rate). The error classification rate decreased by 3.7% to 8.6%.

Fig. 9: Results of checking of machine-allocated company properties: "online shop"

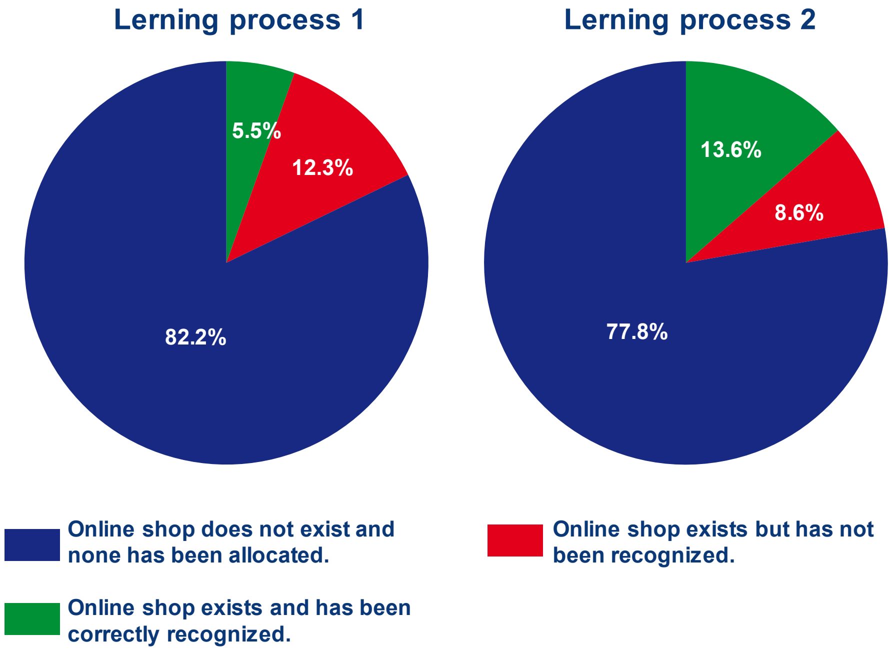
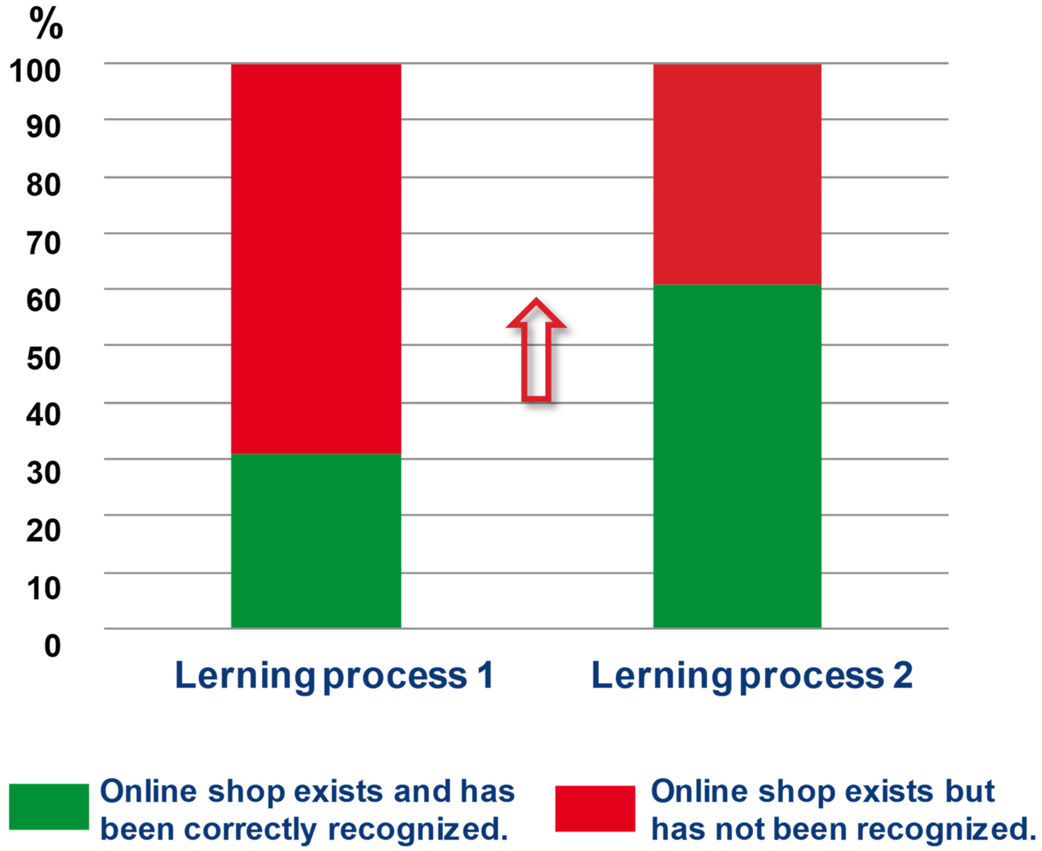


Fig. 10: Proportion of correct classifications of online shops



In the second learning process, the F value rose by 0.288 units to 0.759, which suggests a clearly improved result compared to the first learning process.

The process shows that the learning process should profit from the results in the target database, as the latter becomes more accurate. With continued repetition, the number of addable new datasets from the target dataset should sink to a given level. The process would then come to an end if for any more datasets a probability cannot be predicted that exceeds or falls short of the predefined threshold values.

Assessment of beds available at hotels in the HRS portal

For a methodical project the number of beds available in hotels in the Munich area was required and the publicly accessible portal hrs.de was selected as the source of companies.

On this portal, the following information can be read and is freely accessible:

No. of single rooms, no. of double rooms, no. of rooms overall, name, address (street, house no. and postcode), geo-coordinates (longitude and latitude), hotel category (no. of stars), hotel type and rating.

The structure of the secondary websites with confirmation on the corresponding hotels follows a template similar to the websites obtained via a Google search.

Similar to the procedure that follows the Google results, the source texts are analysed hierarchically until the outcome source texts/hotel site source texts have been found. Using text mining methods, these texts are searched in a context-dependent way for expressions of the relevant features so that these can be saved in a structured way.

In this way, all Munich hotels listed on the HRS portal on July 6 2018 were read and the data on their available beds saved. The dataset contains statistically quantifiable information (expressions of all the features described above).

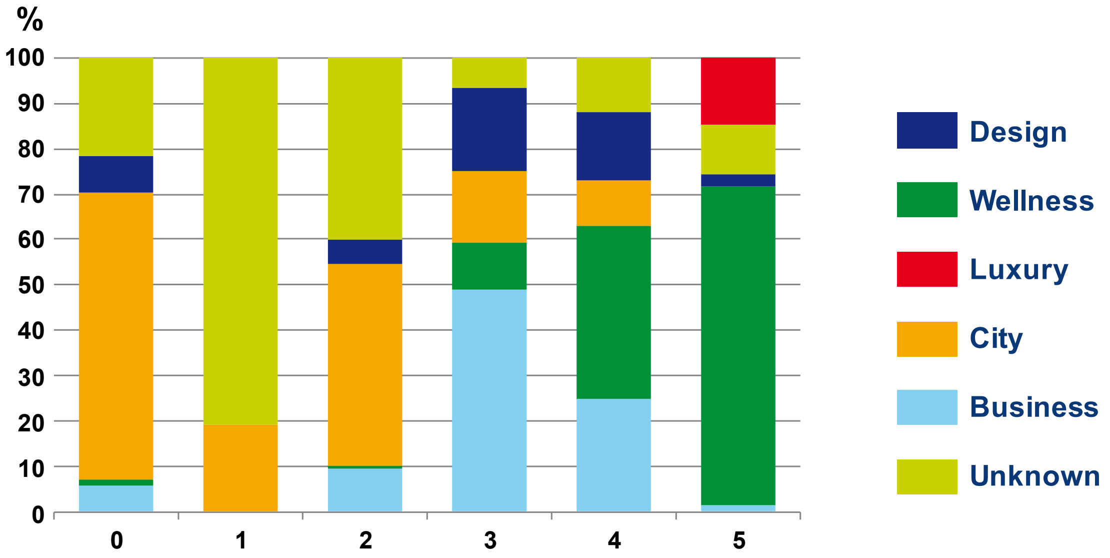
On the day in question, 562 beds were available in Munich. 54.0% of them were double rooms. 25.0% of them were single rooms and 21.0% were other non-defined variations.

Other statistically useful microdata was also available on the portal, such as hotel category and type. Fig. 10 therefore shows the distribution of the 37,562 beds available in the Munich area across the features referred to.

The results show that in Munich most hotels are wellness or city hotels. The relative significance of city hotels decreases as the hotel category rises (star rating) and vice versa for wellness hotels. The latter appear only from three-star category upwards and, at 70.5%, make up the largest proportion of Munich five-star hotels on the portal.

Representing 14.6% of the beds available, luxury hotels are found solely in the highest star rating. Business hotels dominate with 48.9% of the beds on offer in the medium three-star hotel category. At 80.8%, hotels without information on hotel type are mostly in the one-star category.

Fig. 11: Distribution of beds across hotel categories on the HRS portal according to no. of stars and hotel type



At 63.2%, city hotels were the most common.

In the category of hotels that do not participate in the international hotel categorisation process, city hotels were the most represented at 63.2%.

The web scraping of online portals shows that it is well suited to the electronic gathering of comprehensive information about companies in a specific sector without having to use the official register.

It also shows that the assessment of company properties – such as the (in this case) interesting number of beds depending on availability – can be effected electronically relatively easily and accurately.

Conclusion and potential for improvement

The results of the first attempt at web scraping (web scraping + predictive modelling) show that the potential for the classification of company websites based on accuracy has worked very well. The results largely suggest a significant potential for web scraping.

Where predictive modelling is concerned, accuracy essentially depends on the extent and compilation of the training dataset and on the selection of the keywords. This requires a lot of preparatory work in order to be able to obtain high quality training datasets. The procedure has shown that a dynamically designed partly automated strengthened dataset can significantly increase the accuracy of a partly-monitored learning process. The predictive modelling process conducted at HSL via two learning processes will in the future be developed for an unlimited number of repetitions in order to achieve optimal accuracy in binary classification.

The type of keywords, the size of the training database and the predictors for machine learning are strongly dependent on the company property to be predicted. A manually researched training database for the subject areas of e-commerce, conjoined with specifically collected features, cannot be used to automatically assess, for example, companies with green technology or freelancers in the creative sector.

Depending on theme interests and specialist statistical questioning, it will, for the assessment of company characteristics using machine learning methods, for the time being be necessary to issue and validate training datasets with the help of manual research and to maintain them. The online method used in HSL for the automated enrichment of the training data was very helpful in this regard and shows, i.a. with regard to the involvement of modelling diagnostics for machine predictor selection, potential for development.

The technical and methodological implementation of a process of machine learning for predictive identification of latent company properties worked well. Here the use of an R environment is very amenable and sufficiently fast.

Due to the consistent structure of webpages within portals, the search and saving of company websites in online portals can be carried out simply and accurately. Commercial portals may need to be manually researched, but these can be saved in a database and accessed as necessary according to market sector. The scraping of online portals without the use of a meta-search engine has proven to be very fast. The assessment of all Munich hotels listed on the HRS portal took around five minutes using the R algorithm. Not least because of the independence from search engines such as Google, there is no significant potential for improvement immediately apparent here. Further investigation will show to what extent within an overall algorithm already-known portals can be searched and linked and then web scraping with meta-search engines applied to the not-yet-linked company datasets.

For the web scraping of company websites the picture is somewhat different. Given the current capacity limits, the full capture of company-related websites of the approximately 300,000 Hessian companies on the corporate register would take 8.3 years. The development of web scraping will therefore be strongly focused on the increase of capacities in order to be able to implement the full capture of the Hessian company register via web scraping within around 30 days. This would be a satisfactory timeframe.

Bibliography

BARCAROLI, Giulio, Monica SCANNAPIECO and Summa DONATO, 2016. On the Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web. In: *Italian Review of Economics, Demography and Statistics* [online]. **70**(4), p. 25-41. [Accessed on: 03.09.2018]. RePEc. ISSN: 0034-6535, available at: http://www.sieds.it/listing/RePEc/journl/2016LXX_N4_RIEDS_25-41_Scannapieco.pdf

BOTTOU, Leon, and Yann LE CUN, 2004. Large Scale Online Learning. In: *Proceedings from the conference, Neural Information Processing Systems 2003*. Vancouver and Whistler, British Columbia. December 8-13, 2003

BRUNNER, Karola, 2014. Automatisierte Preiserhebung im Internet. In: *Wirtschaft und Statistik.* **4**(2014), p. 258-262. ISSN 1619-2907

CAVALLO, Alberto, 2013. Online and official price indexes: Measuring Argentina's inflation. In: *Journal of Monetary Economics* [online], **60**(2), p. 62-512. [Accessed on: 01.09.2018]. Science Direct. ISSN: 0304-3932. Available at: DOI: 10.1016/j.jmoneco.2012.10.002

COHEN, William W., Pradeep RAVIKUMAR and Stephen E. FIENBERG, 2003. A Comparison of String Metrics for Matching Names and Records. In: *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*. Washington DC, August, 2003

VARGIU, Eloisa., and Mirko URRU, 2013. Exploiting web scraping in a collaborative filtering-based approach to web advertising. In: *Artificial Intelligence Research* [online]. **2**(1), S. 44-54. [Accessed on: 01.09.2018]. Sciedu. ISSN: 1927-6982. Available at: DOI: 10.5430/air.v2n1p44

DREISEITL, Stephan and Lucila OHNO-MACHADO, 2002. Logistic regression and artificial neural network classification models: a methodology review. In: *Journal of Biomedical Informatics* [online], **35**(2002), p. 352-359. [Accessed on: 06.09.2018]. Science Direct. ISSN: 1532-0464. Available at: https://doi.org/10.1016/S1532-0464(03)00034-0

FREES, Edward W., Richard A. DERRIG and Glenn MEYERS, 2014. *Predictive Modelling Applications in the Actuarial Science – Volume I: Predictive Modeling Techniques*. New York: Cambridge University Press. ISBN: 9781107029873

HACKL, P, 2016. Big Data: What can official statistics expect?. In: *Statistical Journal of the IAOS* [online]. **32**(1), S. 43-52 [Accessed on: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Available at: DOI: 10.3233/SJI-160965

HOEKSTRA, Rutger, Olav TEN BOSCH and Frank HARTEVELD, 2012. Automated data collection from web sources for official statistics: First experiences. In: *Statistical Journal of the IAOS* [online]. **28**(3,4), p. 99-111 [Accessed on: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Available at: DOI: 10.3233/SJI-2012-0750

LONG, J. Scott, 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, London, New Deli: SAGE Publications. ISBN: 0803973748

MUNZERT, Simon., Christina RUBBA, Peter MEISZNER, and Dominic NYUIS, 2015. *Automated Data Collection in R: A practical Guide to Web Scraping and Text mining*. United Kingdom: John Wiley & Sons Ltd. ISBN: 111883481X

NILSSON, Nils J.,1998. Introduction to Machine Learning: *An early draft of a proposed Textbook* [unpublished]. Stanford, Stanford University. [Accessed on: 04.09.2018]. Available at: http://robotics.stanford.edu/people/nilsson/mlbook.html

OOSTROM, Lotte, Adam N. WALKER, Bart STAATS, Magda SLOOTBEEK-VAN LAAR, Shirley O. AZURDUY and Bastiaan ROOIJAKKERS, (2016). Measuring the internet economy in The Netherlands: A big data analysis. CBS Discussion Paper 2016/14

POLIDORO, Federico, Riccardo GIANNINI, Rosanna LO CONTE, Stefano MOSCA and Francesca ROSSETTI, 2015. Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. In: *Statistical Journal of the IAOS* [online]. **31**(2), p. 165-176 [Accessed on: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Available at: DOI: 10.3233/sji-150901

SIRISURIYA, SCM de S, 2015. A Comparative Study on Web Scraping. In: *Proceedings of 8th International Research Conference*. General Sir John Kotelawala Defence University, 2015. KDU, p. 135-140

SCHÄFER, Dieter und Matthias BIEG, 2016. Auswirkung der Digitalisierung auf die Preisstatistik. Destatis Methodenpapier. Wiesbaden, Statistisches Bundesamt

Stateva, G., Bosch, O. t., Maślankowski, J., Summa, D., Scannapieco, M., Barcaroli, G., . . . Wu, D. (2017). Work Package 2 – Web scraping Enterprise Characteristics. ESSnet, p. 22.

STATISTISCHES BUNDESAMT, 2015. *Unternehmen und Betriebe im Unternehmensregister: Methodische Grundlagen, Definitionen und Qualität des statistischen Unternehmensregisters*. Statistisches Bundesamt. [Accessed on: 05.09.2018]. Available at: https://www.destatis.de/DE/ZahlenFakten/GesamtwirtschaftUmwelt/UnternehmenHandwerk/Unternehmensregister/Methoden/Methodisches.html

STATISTISCHES BUNDESAMT, 2017. Nutzung von Informations- und Kommunikationstechnologien in Unternehmen 2017. Wiesbaden: Statistisches Bundesamt

TUZHILIN, Alexander, Michele GORGOGLIONE, and Cosimo PALMISANO, 2008. Using Context to Improve Predictive Modelling of Customers in Personalization Applications. In: *IEEE Transactions on Knowledge and Data Engineering* [online], **20**(11), p. 1535-1549. [Accessed on: 05.09.2018]. ISSN 1041-4347. Available at: http://doi.ieeecomputersociety.org/10.1109/TKDE.2008.110

US AIR FORCE, 2006. *Method and Apparatus for improved Web Scraping*. Inventor: SALERNO, John und Douglas M. BOULWARE. 04.07.2006. Application: 26.08.2004. US, Patent spec. US7072890B2

ZWICK, Markus and Lara WIENGARTEN, 2017. Neue digitale Daten in der amtlichen Statistik. In: *Wirtschaft und Statistik*, **5**(2017), p. 19-30. ISSN 1619-2907

Imprint

Copyright:
Hessisches Statistisches Landesamt,
Wiesbaden, 2018

Editor:
Hessisches Statistisches Landesamt,
Wiesbaden, Rheinstraße 35/37
Germany
Telephone: +49 611 3802-0,
Telefax: +49 611 3802-890
E-mail: info@statistik.hessen.de
Website: https://statistik.hessen.de

Contact:
Normen Peters
Telephone: +49 611 3802-517
E-mail: normen.peters@statistik.hessen.de