STATISTIK HESSEN

HESSEN

Webscraping von Unternehmenswebseiten und maschinelles Lernen zum Gewinnen von neuen digitalen Daten

Fachartikel

**Referat PA Grundsatzfragen,
Querschnittsanalysen,
Forschungsdatenzentrum**

Inhaltsverzeichnis

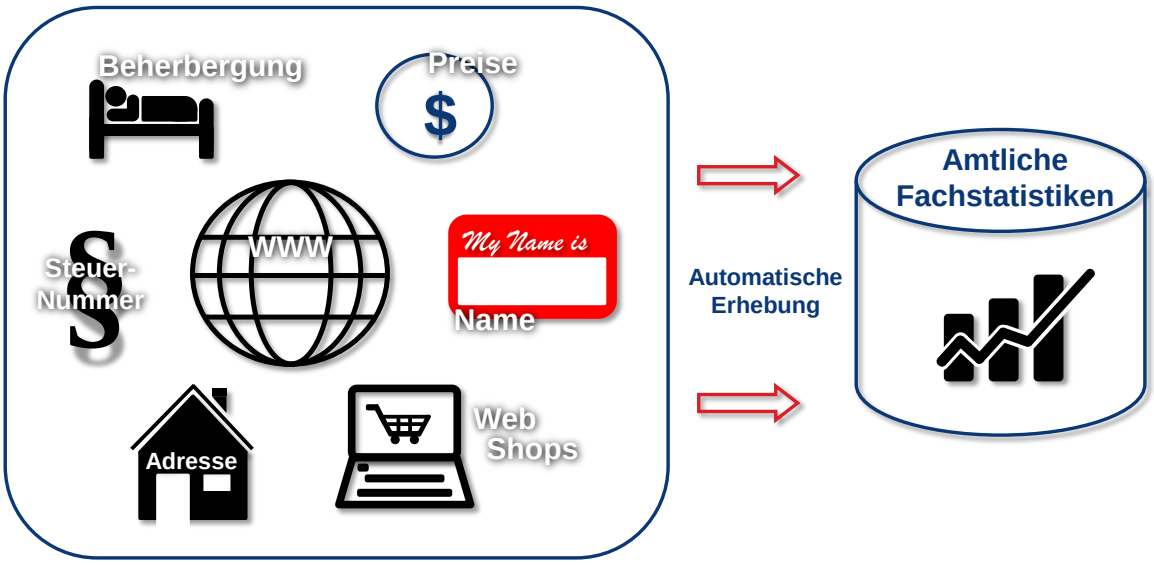
Neue digitale Daten aus dem Internet

Unternehmensdaten aus dem Internet

2017 waren 72 % der deutschen Unternehmen auf einer Webseite präsent, 46 % waren in den sozialen Medien engagiert und 23 % haben Waren oder Dienstleistungen über das Internet vertrieben (siehe Statistisches Bundesamt, 2017). Viele Unternehmensdaten, die durch Befragungen, Interviews oder manuelle Recherchen aufwendig erhoben, plausibilisiert und gepflegt werden, sind daher bereits im Internet vorhanden und öffentlich zugänglich. Die Notwendigkeit, diese Daten durch die Befragung von Unternehmen zu erheben, könnte durch den Einsatz von Webscraping entfallen.

Neue digitale Unternehmensdaten aus dem Internet sind auf kommerziell genutzten Webseiten enthalten, welche Umsätze durch Waren- und Dienstleistungsverkäufe generieren. Dies trifft bspw. auf elektronische Bestell- und Zahlungssysteme wie Online-Shops, kommerzielle Online-Buchungs-Portale oder auf Onlinedienstleistungsanbieter zu. Darüber hinaus sind Unternehmensinformationen über Online-Präsenzen buchbarer Dienstleistungsanbieter z. B. Handwerker oder auf Webseiten zu Informations- oder Marketingzwecken z. B. Friseurläden verfügbar. In beiden Gruppen, welche den Kern der Internetökonomie darstellen, entstehen die Daten prozessbasiert als Nebenprodukte und sind in der Regel nicht Teil einer beabsichtigten Datenproduktion (siehe Oostrom, Walker, Slootbeek-Van Laar, Azurduy und Rooijakkers, 2016).

Abbildung 1: Neue digitale Daten für die Amtlichen Fachstatistiken aus dem Internet



Die neuen digitalen Daten können von verschiedenster Art sein und inhaltlich sehr breite Themengebiete abdecken. Hierzu gehören Daten, die für die Verknüpfung mit amtlichen Daten wichtige Entitäten oder Stammdaten wie Name, Adresse oder Steuernummer aber auch wichtige Fachinformationen wie z. B. Preise von Einzelhandelsunternehmen, Anzahl angebotener Schlafgelegenheiten von Beherbergungsbetrieben oder das Vorhandensein von Onlineshops in der Branche des elektronischen Versandhandels enthalten. Einmal gefunden und öffentlich zugänglich, müssen die Neuen digitalen Daten nur elektronisch erfasst und automatisch erhoben werden.

Vorteile durch das Gewinnen von Internetdaten

Durch das automatisierte Gewinnen der neuen digitalen Daten kann die Amtliche Statistik auf verschiedene Weisen verbessert werden:

- **Erhebungsunterstützung**: Statistische Informationen, die aktuell noch per Befragung, Interview oder Umfragen aufwändig erhoben werden müssen, sind bereits im Internet vorhanden und könnten zukünftig durch Webscraping automatisch und regelmäßig gewonnen werden. Routinearbeiten und manuelle Recherchen würden entfallen. Dies würde den Erhebungsaufwand deutlich reduzieren. Beispiele hierfür sind die Preisermittlung im Rahmen des Verbraucherpreisindexes, die Ermittlung des Bettenangebotes von Beherbergungsbetrieben oder die E-Commerce-Aktivitäten hessischer Handelsunternehmen.

- **Bestandspflegeunterstützung**: Da das Webscraping statistisch quantifizierbare Informationen über Unternehmen liefert, ließen sich diese zur Plausibilisierung verwenden.

- **Schnelligkeit der Datenbereitstellung und Aktualität**: Je nach technischer Kapazität könnte das Webscraping Daten in monatlichen Zyklen liefern.

- **Merkmalskranzerweiterung**: Neben der Unterstützung von Erhebungsverfahren könnte das Webscraping statistisch auswertbare Informationen liefern, die noch nicht in der Amtlichen Statistik vorhanden sind. Dazu gehören zum Beispiel die Beteiligung von Unternehmen an E-Commerce-Aktivitäten, Investitionsaktivitäten in nachhaltige Technologie, Sicherheitsstandards auf Unternehmenswebseiten oder die Barrierefreiheit auf Internetshoppingseiten.

- **Entlastung von Auskunftspflichtigen**: Öffentlich zugängliche Daten müssten von den Auskunftspflichtigen nicht im Rahmen der Berichtspflicht an die Amtliche Statistik gemeldet werden. Eine Entlastung der Auskunftspflichtigen wäre möglich.

Zusammenfassend lässt sich vom automatisierten Extrahieren digitaler Daten aus dem Internet erwarten, dass die Produkte der Amtlichen Statistik in kürzerer Zeit, zu geringeren Kosten, mit geringerem Befragungsaufwand, mit detaillierterer Aufgliederung und umfänglicher inhaltlicher Diversität verfügbar gemacht werden könnten. Ein vollständiges Ersetzen des herkömmlichen Statistikherstellungsprozesses ist nach den bisherigen Erfahrungen in naher Zukunft nicht zu erwarten, sondern eher eine umfangreiche Ergänzung der bisherigen Prozeduren (siehe Hackl, 2016).

Webscraping – Hintergrund

Suchen, Finden, Strukturieren und Speichern von Daten

Das Internet wird oft als große Bibliothek von miteinander verbundenen Ressourcen verstanden. Unter den Ressourcen befinden sich die interessierenden Daten. Das Problem besteht darin diese vollständig in einem angemessenen Zeitrahmen und mit vertretbarem Aufwand zu finden. In diesem Zusammenhang wurden Metasuchmaschinen populär, die

dadurch charakterisiert sind, dass Suchanfragen an mehrere Suchmaschinen gleichzeitig weitergeleitet und die Ergebnisse aufbereitet werden können.

Heutzutage nutzen viele Metasuchmaschinen Webscraping. Dieses Verfahren umfasst allgemein Prozesse, die Entitäten aus Quelldatenbanken abfragen, an Metasuchmaschinen weiterleiten und so das Finden der gesuchten Webseiten ermöglichen. Die gefundenen Webseiten werden anschließend nach den entsprechenden Inhalten durchsucht, welche dann mittels Informationsextraktion gewonnen, transformiert und mit anderen Datenbanken verknüpft werden (siehe Salerno und Boulware, 2006).

Somit wird das Webscraping vorrangig mit dem Ziel eingesetzt, unstrukturierte Informationen auf Internetseiten zu finden, zu extrahieren, diese in verständliche Formate zu strukturieren und somit für Datenbanken, Tabellen oder kommaseparierte Textdateien speicherfähig zu machen (siehe Sirisuriya, 2015).

Webscraping auf europäischer und internationaler Ebene

Durch den Anstieg von online getriebenem Handel und online getriebener Kommunikation hat das Webscraping-Verfahren in der Amtlichen Statistik der nationalen Statistikämter bereits Anwendung mit guten Ergebnissen gefunden. In einer der frühen Machbarkeitsstudien des automatisierten Gewinnens von digitalen Daten für das Niederländische Statistische Amt (CBS) ergab sich bspw., dass das Erheben und Weiterverarbeiten von digitalen Daten mit Webscraping möglich ist, deutliche Effizienz- und Lerneffekte erzielt werden und insbesondere bei großen Datenmengen zu Vorteilen bzgl. Datenbereitstellungsgeschwindigkeit und Datenqualität führt. Dabei müssen jedoch insbesondere die Kosten, die durch Anpassungen der Verfahren bei Änderungen der Webseiteninfrastruktur anfallen, berücksichtigt werden (siehe Hoekstra, ten Bosch und Harteveld, 2012).

Ein erster häufiger Anwendungsbereich für das Webscraping nationaler amtlicher Statistikinstitute war das automatisierte Erheben von Konsumentenpreisen. Das Verfahren wurde bspw. erfolgreich für das Berechnen der argentinischen Online-Inflationsrate mit Daten von Onlinehändlern von 2007 – 2011 genutzt. Die Online-Inflationsrate übertraf dabei die herkömmlich berechnete um das Dreifache (siehe Cavallo, 2013). Auf europäischer Ebene hatte sich das italienische nationale Statistikamt (ISTAT) erfolgreich mit der automatisierten Erhebung von Konsumentenpreisen im Internet per Webscraping innerhalb des europäischen Projekts „Multipurpose Price Statistics (MPS)“ beteiligt (siehe Polidoro und andere, 2015). Das Statistische Bundesamt verwendet das Webscraping bereits seit einigen Jahren erfolgreich und in zunehmendem Umfang in seiner Preisstatistik (siehe Brunner, 2014 oder Schäfer und Bieg, 2016). Nachfolgend wurde das Webscraping auch auf andere Bereiche der nationalen Statistiken ausgedehnt.

EUROSTAT und nationale, statistikbezogene Behörden und Institute gründeten das Netzwerk des europäischen statistischen Systems (ESSnet), um auf europäischer Ebene vergleichbare Statistiken zu produzieren. Innerhalb des Netzwerkes wurde das ESSnet-Projekt „Big Data“ nach einer Ausschreibung der europäischen Kommission von 22 nationalen Partnern beschlossen und ins Leben gerufen. Dieses Projekt hatte die Integration von Big Data in die europäischen amtlichen Statistiken zum Ziel. Es bestand aus insgesamt 8 Arbeitspaketen, die das Gewinnen von Neuen Digitalen Daten über verschiedene Methoden und Wege beinhalteten. Die Arbeitspakete 1 und 2 deckten dabei das Ermitteln von neuen digitalen Daten mit Webscraping-Verfahren ab.

Das Arbeitspaket WP1 „Webscraping job vacancies“ befasste sich mit dem automatischen Extrahieren von Informationen über Jobangebote auf u. a. Jobportalen oder Unternehmenswebseiten. Neben den Ländern Tschechien, Italien, Großbritannien und Irland, hatte sich das Statistische Bundesamt für Deutschland mit einem eigenen Pilotprojekt innerhalb dieses europäischen Rahmens beteiligt: Eine Machbarkeitsstudie zur Erfassung von Stellenausschreibungen auf Jobbörsen (GigaJob.de, Online-Stellenmarkt.net, Jobs.meinetadt.de) für den deutschen Arbeitsmarkt (siehe Zwick und Wiengarten, 2017).

Das Arbeitspaket WP2 „Webscraping enterprise characteristics“ thematisierte die automatisierte Suche, Speicherung, Strukturierung und Verknüpfung von Unternehmenswebseiten mit den Datensätzen der amtlichen Fachstatistiken. Das Ziel war, bestehende Wirtschafts- und Unternehmensregister mit den digitalen Unternehmensinformationen anzureichern und zu verbessern. Im Rahmen des Projektes wurden als experimentelle Statistiken folgende Merkmale von Unternehmen auf Basis der nationalen Unternehmensregister erhoben:

- Anzahl von Unternehmenswebseiten,
- Unternehmen mit E-Commerce-Aktivitäten,
- Anzahl von Stellenangeboten auf Unternehmenswebseiten,
- Präsenz der Unternehmen in den sozialen Medien.

Hier waren die nationalen Statistikinstitute der folgenden Länder beteiligt: Italien, Bulgarien, Niederlande, Polen, Großbritannien und Schweden. Deutschland war an diesem Projekt nicht beteiligt. Das Italienische Nationale Statistikamt (ISTAT) war hierbei federführend. ISTAT entwickelte eigene Java-Such-Routinen und hat mit 78 000 Unternehmenswebseiten mit Abstand den größten Beitrag an der automatisierten Informationsextraktion von Unternehmenswebseiten geleistet. Die Suchroutinen wurden auch von den beteiligten Ländern Polen und Bulgarien erfolgreich angewandt.

Die in den Arbeitspaketen enthaltenen Machbarkeitsstudien wurden von Februar 2016 bis Mai 2018 durchgeführt und mit teils guten Ergebnissen zum Webscraping von Unternehmenswebseiten und Unternehmenseigenschaften fertiggestellt. In den Pilotprojekten kamen die sechs teilnehmenden, nationalen Statistikämter zu dem Schluss, dass mit dem Webscraping-Verfahren mit verschiedenen Methodenansätzen hochwertige Ergebnisse erzielt werden können, die Verfahren jedoch sehr aufwendig sind und noch vor vielen Herausforderungen stehen.

In einer neuen Ausschreibung der Europäischen Kommission für ein weiteres EU-weites Forschungsprojekt („ESSnet Big Data II“, 2018-2020) wurden weitere fünf mögliche Pilotprojekte definiert. Das Arbeitspaket „Smart Tourism“ widmet sich dabei dem Thema „innovative Datenquellen und Methoden in der Tourismusstatistik“. Viele Daten, die für die Tourismusstatistik von Relevanz sind, sind heutzutage ebenfalls im Internet auf Onlineportalen wie u. a. Reiseportalen, Buchungsportalen oder Webseiten von Beherbergungs- und Tourismusbetrieben vorhanden. Sollte dieses Arbeitspaket zu den mindestens drei geförderten Projekten im europäischen Rahmen zählen, könnte das Webscraping als fundamentale Methode zum Extrahieren der digitalen Tourismusdaten aus dem Internet dabei eine zentrale Rolle spielen.

Webscraping im Hessischen Statistischen Landesamt

Das Hessische Statistische Landesamt unternimmt seit Oktober 2017 verschiedene Maßnahmen und Aktivitäten um Webscraping einzusetzen. Wichtige Orientierungshilfen lieferten die Machbarkeitsstudien im Arbeitspaket WP2 „webscraping of entreprise characteristics“ des ESSnet-Projekts „Big Data“ und insbesondere die von ISTAT freundlicherweise bereitgestellten Algorithmen.

Kern der Anwendung war das hessische Unternehmensregister mit den wichtigen Stammdaten der etwa 300 000 in Hessen ansässigen Unternehmen. Ziel war es, die Internetseiten von Unternehmen zu finden, die öffentlich zugänglichen Daten zu verknüpfen und auszuwerten.

Die Algorithmen zum Finden, Auslesen, Strukturieren und Verknüpfen der auf Unternehmenswebseiten vorhandenen Daten wurden auf eine Stichprobe hessischer Unternehmen aus dem Datenbestand der amtlichen Statistik mit guten Ergebnissen angewendet. Wie in den nächsten Punkten im Detail dargestellt, wurden in einem ersten Schritt erfolgreich Verknüpfungen zu Unternehmenswebseiten zahlreicher hessischer Unternehmen erzielt. Im Anschluss daran erfolgte die Auswertung eines Teils der auf den verknüpften Unternehmenswebseiten auslesbaren Schlagwörter mit Methoden des Text Minings. Final wurde mit Methoden der prädiktiven Modellierung mithilfe von Trainingsdaten und einer Eingang-Ausgang-Funktion das Vorhandensein eines Onlineshops für hessische Unternehmen maschinell bestimmt.

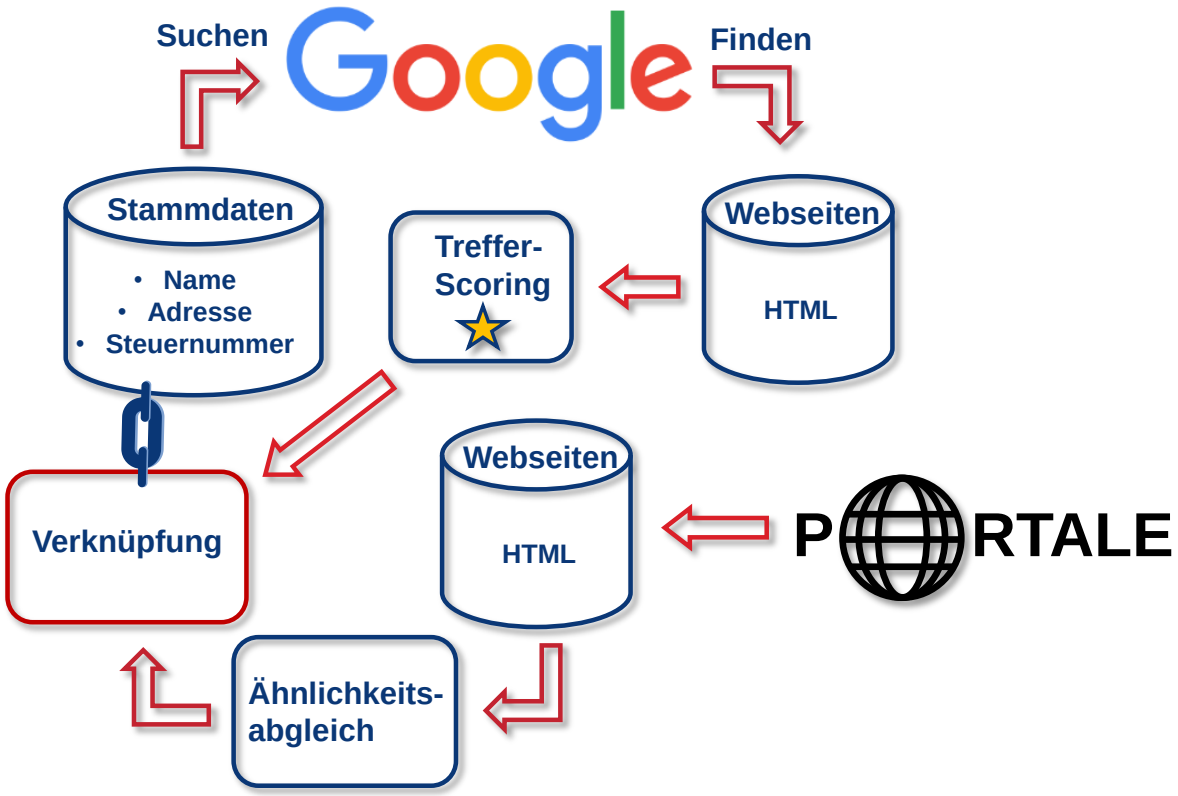
Funktionsweise – Suchen, Finden und Verknüpfen

Die Suche nach quantifizierbaren Inhalten im Internet mit Informationsextraktion wird oft von sogenannten Webcrawlern durchgeführt. Beim Webcrawling werden grundsätzlich alle auffindbaren Internetseiten automatisch und methodologisch gesucht und gespeichert. Beim Webscraping werden nach bestimmten, vordefinierten Informationen gesucht sowie quantifizierbare Inhalte strukturiert und extrahiert. Im vorliegenden Fall wird nach existierenden in Hessen ansässigen Unternehmen gesucht.

Das Webscraping nutzt identifizierende Merkmale oder Vorinformation bei der Suche nach Fachinhalten und Links. Öffentlich zugängliche Daten sind häufig unstrukturiert und müssen in eine für die Auswertung geeignete Form transformiert werden (siehe Vargiu und Urru, 2013).

Da bei diesem Projekt quantifizierbare Informationen von Unternehmenswebseiten extrahiert werden sollten, wurden die registerbasierten, in Datenbanken gespeicherten Stammdaten der Amtlichen Statistik und in kommerziellen Online-Buchungsportalen öffentlich zugänglichen Daten als Vorinformationen betrachtet. Stammdaten, die Unternehmen identifizieren können, enthalten: Adressen, Firmenbezeichnungen, Rechtsformen, Informationen über zugehörige Niederlassungen und unselbständige Zweigstellen. Diese Informationen können datensatzweise für die Suche im Internet genutzt werden. Gleichzeitig eignen sich aus den Datenbanken extrahierbare Datentabellen dafür mit neuen digitalen Informationen angereichert zu werden (siehe Abbildung 2).

Abbildung 2: Verknüpfung von Webseiten mittels Google- und Portalsuche



Scraping von Unternehmenswebseiten mittels Metasuchmaschine

Das vollautomatisierte Webscraping ist dafür geeignet mit verfügbaren Unternehmensinformationen aus dem Datenbestand der Amtlichen Statistik Daten und quantifizierbare Inhalte zu identifizieren, zu speichern, zu strukturieren, auszuwerten und mit den Daten der amtlichen Statistik zu verknüpfen.

Angesichts der umfassenden und erfolgreichen Aktivitäten des italienischen nationalen Statistikamts (ISTAT) bei der automatisierten Informationsextraktion von Unternehmenswebseiten, hat sich das HSL eng an der von ISTAT zur Verfügung gestellten Prozedur orientiert (siehe Barcaroli, Scannapieco, und Donato, 2016).

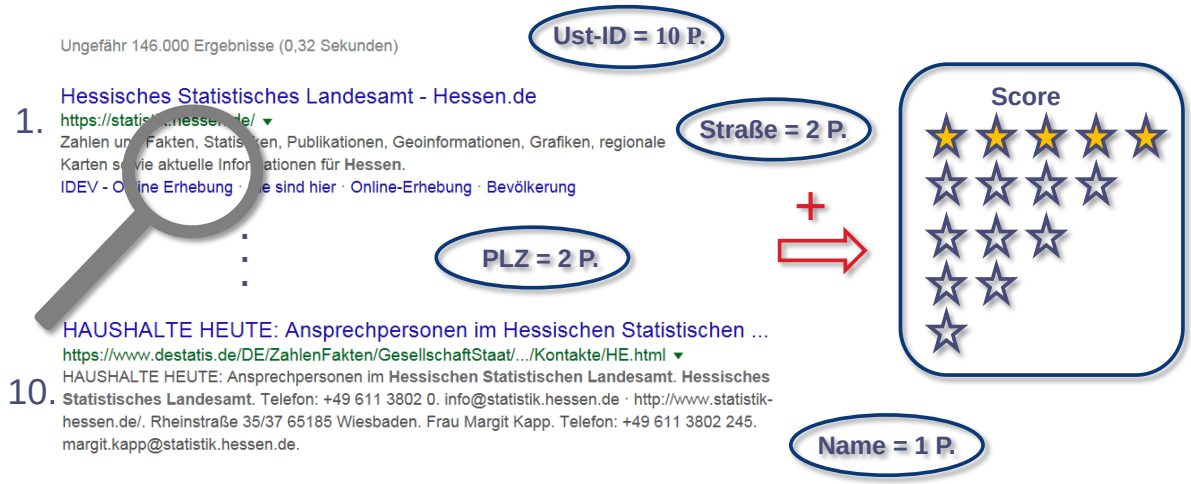
Die anzureichernden Datensätze, welche die Identifikationsmerkmale/Entitäten wie Name, Adresse oder Steuernummer enthalten, werden zunächst in einer Datenbank gespeichert. Ein in der Programmiersprache Java geschriebener Algorithmus greift von dort aus auf die einzelnen Datensätze zu, um diese für die weitere Suche mit einem URL-Crawler zu verwenden. Um die Internetseiten über die Identifikationsmerkmale suchen zu können, muss eine Metasuchmaschine verwendet werden. Hier bietet sich die Metasuchmaschine Google.com an. ISTAT hatte die Suchmaschine Bing genutzt, Google hat sich jedoch für das HSL bei der Suche als treffsicherer erwiesen.

Die identifizierenden Merkmale aus der Amtlichen Statistik werden bei der Suche mit Google nach Unternehmenswebseiten automatisiert und datensatzweise genutzt. Nach Eingabe der Merkmale eines Unternehmens enthält die Ergebnisseite der Google-Suche nun gefundene Internetreferenzen in einer bestimmten Anzahl.

Nun werden die Quelltexte/Webseiten der 10 höchstplatzierten Unternehmenswebseiten (Hauptwebseiten) sowie die darin enthaltenen identifizierenden Sekundärwebseiten (Impressum, Kontakte oder „Über uns“) und das Google-Knowledge-Panel auf der rechten Ergebnisseite (Google-Rechte-Hand-Seite) gespeichert und nach passenden Stammdaten durchsucht. Pro Eingabe/Unternehmen werden also potentiell bis zu 44 Webseiten durchsucht.

Die gespeicherten Haupt- und Sekundärseiten werden nun abhängig von der Anzahl, Art und Ausprägung der gefundenen Identifikationsmerkmale über ein Punktesystem bewertet. Die vergebenen Noten werden über die Sekundärseiten aufsummiert und ergeben ein gewichtetes Treffer-Scoring. Weiterhin werden die Webseiten dann nach Punkteanzahl und der Google-Platzierung geordnet. Anschließend wird die Hauptunternehmensseite mit der höchsten Benotung oder der besten Google-Platzierung den entsprechenden Stammdaten zugeordnet. Die Zweifachsortierung gewährleistet dabei eine eineindeutige Sortierung. Auf diese Weise wird die Anzahl gefundener, zuordnungsfähiger Webseiten pro Unternehmen auf eine reduziert.

Abbildung 3: Bewertungssystem



Auf diese Weise ist es möglich, alle auf den zugeordneten Webseiten öffentlich zugänglichen Informationen, sofern quantifizierbar, zu verknüpfen. Es hätte der Fall eintreten können, dass keine der 11 Ergebniswebseiten der Suchmaschine zuordnungsfähig ist. In diesem Fall wäre die Punktereihenfolge nicht davon berührt, jedoch das Niveau der Gesamtpunktezahl. Dass die zugeordnete Webseite zunächst „richtig“ war, wurde bei ISTAT an dieser Stelle automatisiert mit einem auf Wahrscheinlichkeiten beruhendem in R programmierten Machine Learning Verfahren sichergestellt. Im HSL hat sich ein regelbasierter Ansatz als ausreichend herausgestellt. Dieser beinhaltete das Definieren einer zu erreichenden Mindestpunktezahl von 5 als erste, „initiale“ Wahrheitsüberprüfung.

Scraping von kommerziellen Onlineportalen

Viele Kleinunternehmen sind oft nicht auf einer eigenen Webseite, sondern auf einem kommerziellen Onlineportal präsent. Das automatisierte Finden, Speichern, Strukturieren und Verknüpfen von Daten aus Onlineportalen funktioniert in dieser Untersuchung anders als das Webscraping mittels Meta-Suchmaschine. Das von ISTAT entwickelte Verfahren versucht bspw. die passende Impressums-Webseite des jeweiligen Unternehmens herunterzuladen, um darauf identifizierende Merkmale finden zu können. Bei kommerziellen Online-Portalen

wird man auf diese Weise nur den Portalbetreiber finden, nicht jedoch das interessierende Unternehmen. Deshalb wurde im HSL in einem Feldversuch ein eigener Algorithmus zur Informationsextraktion aus einem kommerziellen Onlinebuchungsportal entwickelt und programmiert.

Onlineportale können als fachthemenabhängige (bspw. Jobportale, Beherbergungsportale oder Immobilienportale), kommerziell genutzte Verzeichnisse/Zusammenstellungen verschiedener Unternehmen in ihrer Funktion als Anbieter eines bestimmten Produktes, einer bestimmten Branche, in verschiedenen Regionen betrachtet werden. In der Regel wünschen die Unternehmen, von potentiellen Kunden auf diesen Onlineportalen gefunden und beauftragt zu werden. Insbesondere Kleinunternehmen oder Freiberufler ohne eigene Webseite sind auf solchen Onlineportalen registriert.

Die Kenntnis über den Wirtschaftszweig bzw. über die Branche von Unternehmen ermöglicht es also, vor, parallel oder nach dem Webscraping-Verknüpfungs-Prozess, Onlineportale nach Unternehmen automatisch abzusuchen. Das Onlineportal wird nicht nach dem jeweiligen Unternehmen, sondern nach allen Unternehmen des jeweiligen Wirtschaftszweiges, in der jeweiligen Stadt mittels URL-Crawling abgesucht. Einmal gefunden, können die URLs der jeweiligen Portalsubwebseiten mit den Datensätzen der Amtlichen Statistik verknüpft werden.

Das Portalscraping, zusätzlich zum Webscraping durchzuführen, hat dabei folgende zusätzliche Vorteile:

- **Erfassung abseits von Relevanzschwellen**: Durch das Portalscraping können Unternehmen erfasst werden, die aufgrund bestimmter Umsatz- oder Beschäftigtenkonstellationen nicht in den Stammdaten der Amtlichen Statistik enthalten sind. Dies trifft etwa auf Freiberufler oder auf nicht umsatzsteuerpflichtige Kleinunternehmen zu.

- **Effizienz:** Die Anzahl zu durchsuchender Webseiten ist um ein Vielfaches geringer als beim Webscraping mittels Metasuchmaschine, da die Online-Portale ausschließlich Webseiten des jeweiligen kommerziellen Bereichs enthalten. Das Portalscraping benötigt weniger Speicherplatz und hat sich als um ein Vielfaches schneller herausgestellt als das Webscraping von Unternehmenswebseiten.

- **Geringere Komplexität:** Die Anzahl der Arbeitsschritte sind beim Portalscraping ebenfalls geringer, was das Portalscraping einfacher macht als das Webscraping von Unternehmenswebseiten. Die Unternehmenseigenschaften und Stammdaten sind auf den verschiedenen Subwebseiten des Onlineportals immer gleich strukturiert. Daher können die Suchalgorithmen einfacher programmiert werden.

- **Genauigkeit:** Das Auslesen und Zuweisen der digitalen Informationen von Online-Portalen hat aufgrund des Bezugs zur gesuchten Branche und Region eine sehr hohe Treffergenauigkeit. Deshalb muss eine Bewertung hinsichtlich der Trefferqualität nicht durchgeführt werden. Zur Verknüpfung mit den Stammdaten der Amtlichen Statistik muss lediglich ein Ähnlichkeitsabgleich durchgeführt werden.

- **Datensparsamkeit:** Die automatisierte Eingabe der Stammdaten in eine Suchmaschine ist für das Portalscraping nicht erforderlich und wird erst bei der Verknüpfung benötigt. Die Erhebung und Verarbeitung/Auswertung der digitalen Daten in Portalen ist jedoch schon vor der Verknüpfung möglich. Beim Webscraping können Suchen

nach digitalen Unternehmensinformationen ohne vorhandene Stammdaten nur schwierig durchgeführt werden.

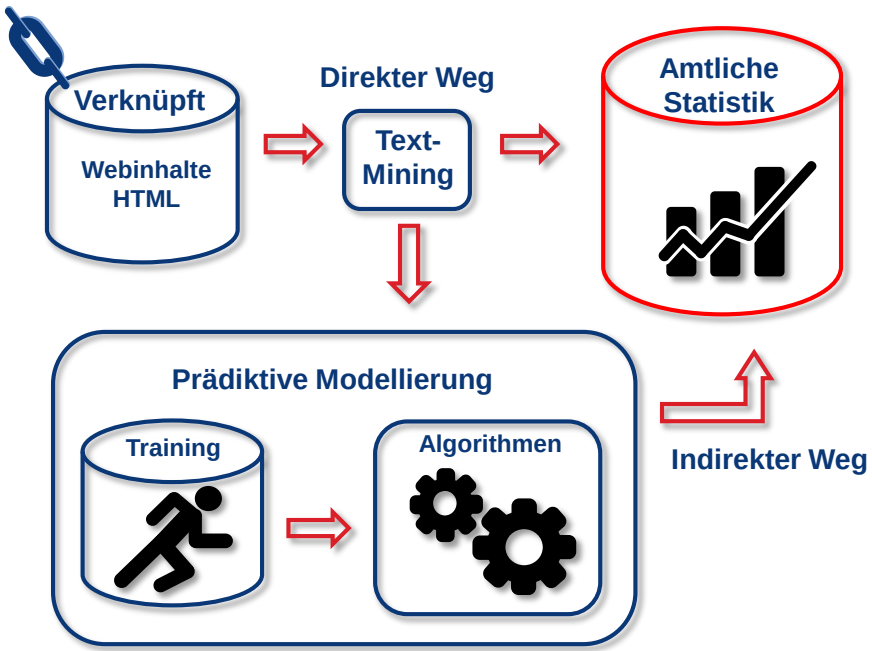
Damit das Portalscraping durchgeführt werden kann, müssen die passenden Online-Portale je nach Wirtschaftsbranche und Themengebiet vor der automatisierten Erhebung manuell im Internet recherchiert werden. Die Zahl der passenden Online-Portale ist jedoch themenabhängig begrenzt und nachhaltbar. Der Aufwand bei der manuellen Recherche der Online-Portale wird deshalb als vertretbar angenommen.

Für die Verknüpfung der gefundenen Inhalte ist es erforderlich einen geeigneten Ähnlichkeitsabgleich durchzuführen. Hier bieten sich einfach umsetzbare und in R oder Java programmierbare, metrische Wortdistanzmaße an, wie beispielsweise die Levenstein Distanz, die Monge-Elkan-Distanz oder die Jaro-Winkler-Distanz. Diese basieren auf dem einfachen Vergleich von Zeichen und Buchstaben verschiedener Wörter (siehe Cohen, Ravikumar, und Fienberg, 2003).

Datenanreicherung mit verknüpften Inhalten

Wie in Abbildung 4 illustriert, können die durch das Webscraping gewonnenen Daten auf verschiedenen Wegen mit der Amtlichen Statistik verknüpft werden. Dabei gibt es eine direkte Übertragung, in der die gefundenen Daten strukturiert und als Indikatoren, Schlüsselwortzähler oder auslesbare Merkmalsausprägungen in den Datenbestand der amtlichen Statistik gebracht werden.

Abbildung 4: Weg der möglichen Anreicherung der Datensätze durch die Webinhalte



Sind gewünschte Unternehmenseigenschaften nicht durch einfaches Auslesen zu ermitteln, verbleibt noch die Möglichkeit der Datenanreicherung mit Hilfe von Verfahren des maschinellen Lernens (indirekter Weg).

Direkte Datenanreicherung

Wenn die entsprechenden Webseiten (Unternehmenswebseiten und unternehmensbezogene, sekundäre Onlineportalseiten) mit den statistischen Einheiten (Bestandsdaten) der Amtlichen Statistik verknüpft worden sind, können diese auf fachthemenabhängige Inhalte hin untersucht werden.

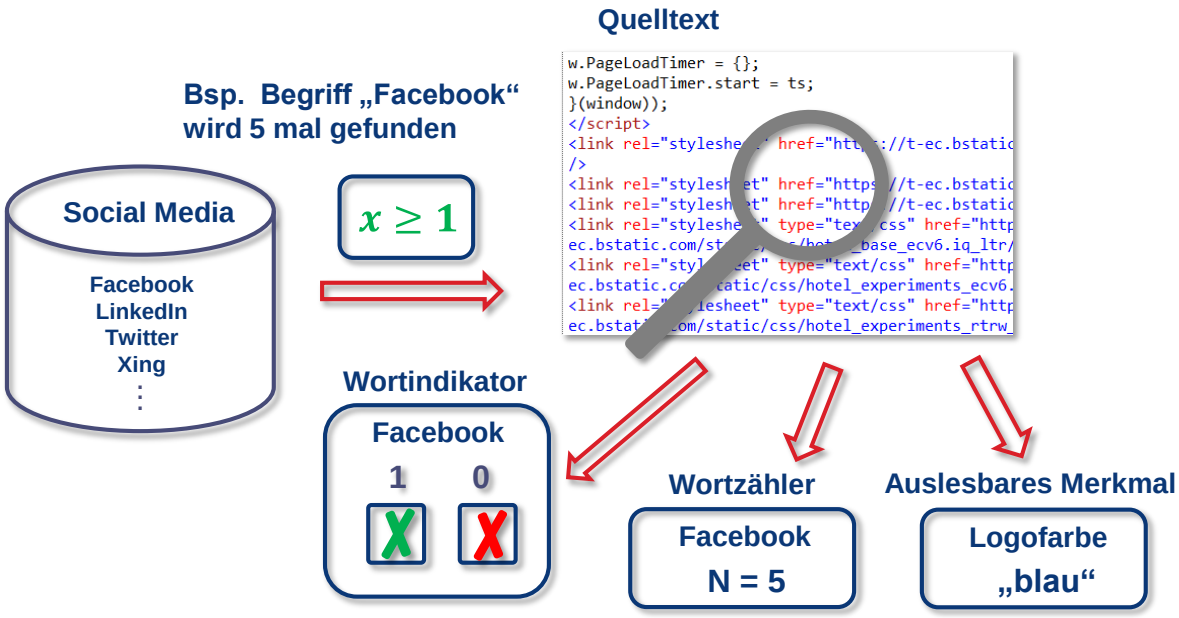
Dazu werden die zugehörigen Quelltexte zunächst heruntergeladen. Die in den Quelltexten enthaltenen Daten sind noch keine quantifizierbaren Informationen, sondern liegen zunächst nur in unstrukturierter oder in unterschiedlich strukturierter Form vor. Die Möglichkeiten, die auf den Unternehmenswebseiten vorliegenden unstrukturierten Daten in quantifizierbare, strukturierte statistische Informationen aufzubereiten sind im Folgenden dargestellt (siehe auch Abbildung 5):

- **Schlüsselwortabhängiges Generieren von Indikatoren:** Die Wortindikatoren bekommen bei Vorkommen des gewünschten Schlüsselwortes den Wert 1 zugewiesen. Ist das entsprechende Wort nicht auf der Webseite zu finden, bekommt der jeweilige Wortindikator den Wert 0 zugewiesen.

- **Schlüsselwortzähler:** Hier wird die Häufigkeit des Vorkommens des gesuchten Schlüsselwortes übergeben.

- **Auslesbare Merkmale:** Sind Merkmalsausprägungen auf der Webseite auslesbar wie bspw. die Anzahl von Doppelzimmer auf der Webseite eines Beherbergungsbetriebs oder die Logofarbe der Webseite „Facebook“, dann kann die Ausprägung in eine entsprechende direkt Variable übergeben werden.

Abbildung 5: Strukturierung unstrukturierter Quelltextdaten



Die Voraussetzung für die direkte Datengewinnung in heruntergeladenen Quelltexten mit den drei beschriebenen Methoden, ist das kontextabhängige Auftreten von Schlüsselwörtern. Bspw. kann das Schlüsselwort „Einzelzimmer“ auf Webseiten von Beherbergungsbetrieben mehrfach in mehreren Kontexten auftauchen, etwa in der Bemerkung eines Hotelgasts in

einer persönlichen Bewertung mit eigenen Worten. Für das automatisierte Erheben der Anzahl von Einzelzimmern in dem jeweiligen Beherbergungsbetrieb ist es wichtig, welche Zeichen, Worte und Muster vor und nach dem gefundenen Schlüsselwort auftauchen.

Hier ist das Nutzen regulärer Ausdrücke als Methode des Text-Minings für die Strukturierung, bspw. in R, unumgänglich (siehe Munzert, Rubba, Meißner und Nyuis, 2015).

Die Daten nach der beschriebenen Weise der Strukturierung in die Datensätze der amtlichen Statistik zu übernehmen, wird als direkter Weg der Datengewinnung über das Internet nach Verknüpfung betrachtet.

Indirekte Datengewinnung – Maschinelles Ermitteln von Unternehmenseigenschaften

Bestimmte Unternehmenseigenschaften, beispielsweise E-Commerce-Aktivitäten sind u. a. durch die Inbetriebnahme von Onlineshops zum elektronischen Vertrieb von Waren und Dienstleistungen charakterisiert. Das Vorhandensein eines Onlineshops kann nicht immer direkt auf einer Webseite ausgelesen werden, obwohl die Webseite einen solchen enthält. Statistische Methoden zur Bestimmung latenter Eigenschaften können helfen, solche Eigenschaften sichtbar zu machen und die enthaltenen Webseiten binär zu klassifizieren. Die Voraussetzung für solche Klassifizierungen ist, dass sich Webseiten, die einen Onlineshop beinhalten, durch verschiedene Merkmalsausprägungen bspw. von reinen Onlinepräsenzen zu Marketingzwecken unterscheiden.

Bestimmte Merkmale, die durch die skizzierte Strukturierung vorliegen, können mit der gewünschten, gesuchten Unternehmenseigenschaft in einer Verbindung stehen. Bei der Eigenschaft „Onlineshop“ kann dies u. a. auf das Vorhandensein eines Zahlungssystems, eines Wareneinkaufssystems, einer Verlinkung zu den sozialen Medien oder durch das Handeln mit bestimmten Waren zutreffen. Sofern eines oder mehrere solcher Merkmale auf einer Webseite vorkommen, ist es möglich, diese Internetseite als einen Onlineshop enthaltend zu klassifizieren. Diese Verfahrensweise kann als „Unternehmenseigenschaftszuordnung nach deterministischen Entscheidungsregeln“ beschrieben werden.

Der Nachteil an dieser Methode ist, dass der Zusammenhang zwischen den Merkmalen und der gesuchten Eigenschaft bekannt sein muss und die Entscheidungsregel innerhalb der Suche nicht verändert oder angepasst werden kann. Es muss vorentschieden werden, wie viele Merkmale auf welche Weise bestimmte Eigenschaften eines Onlineshops determinieren. Es ist jedoch dann möglich, dass die beschriebenen Merkmale auch auf Internetseiten vorkommen, die nicht als Onlineshops enthaltend klassifiziert werden können. In diesem Fall wäre die Verwendung einer deterministischen Entscheidungsregel zu fehleranfällig.

Prädiktive Modellierung

Die Prädiktive Modellierung ist ein beliebtes Verfahren des maschinellen Lernens, welches es ermöglicht, die geschätzte Kausalität zwischen den erhobenen Merkmalen und der Wahrscheinlichkeit für das Auftreten der gewünschten Unternehmensinformation als Entscheidungsgrundlage zu berücksichtigen. Häufige Anwendungsbereiche sind u. a. die Versicherungswirtschaft und Business Intelligence: Hier werden die Algorithmen der prädiktiven Modellierung genutzt um Kunden zu segmentieren, Umsätze zu prognostizieren, Märkte zu analysieren oder um Risiken einzuschätzen (siehe Frees, Derrig und Meyers, 2014).

Doch auch für Onlinemarketing, Spamidentifikation, Betrugsprävention oder für das Customer Relationship Management (Identifizieren und Segmentieren von potentiellen Kunden nach Kaufwahrscheinlichkeit) wird dieses Verfahren häufig verwendet. Mit Hilfe von historischen Daten kann dabei festgestellt werden, welche Produktarten die Benutzer interessieren könnten oder auf welche Felder, Buttons und Links sie wahrscheinlich klicken (siehe Tuzhilin, Gorgoglione und Palmisano, 2008).

Funktionslernen

Die Verfahren der prädiktiven Modellierung beruhen auf Wahrscheinlichkeiten des Auftretens der interessierenden Eigenschaft, dessen kausale Beziehung zu den erhobenen Merkmalen durch eine unbekannte Funktion f dargestellt wird. Die interessierende Eigenschaft dieser Funktion ist nominal skaliert, hat somit einen booleschen Ausgangswert und wird als Klassifizierer bezeichnet. Der boolesche Wert besteht aus einem positiven Fall wenn die interessierende Eigenschaft aufritt (Bspw. Onlineshop) und einem negativen Fall wenn diese nicht auftritt. Eine hypothetische Eingang-Ausgang-Funktion h wird nun definiert. Die Form dieser Funktion ist beliebig und folgt hier einer logistischen Verteilung.

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

Die Parameter β stehen dabei für die kausale Beziehung zwischen den Erhebungsmerkmalen in X und den Auftrittswahrscheinlichkeiten. Der Ausgang der hypothetischen Funktion ergibt dabei die geschätzte Wahrscheinlichkeit für das Auftreten der interessierenden Eigenschaft. Somit liegt dem hier genutzten Methodenansatz innerhalb des Lernalgorithmus eine logistische Regression zugrunde (siehe Long, 1997).

Es gibt verschiedene Methoden der prädiktiven Identifikation, die Logistische Regression ist jedoch eine sehr populäre und leicht nachvollziehbare Methode, die sich insbesondere im Bereich „pattern recognition“ in der Medizininformatik bewährt hat, auf Wahrscheinlichkeiten beruht und einfach anzuwenden ist (siehe Dreiseitl und Ohno-Machado, 2002). ISTAT hat bei der Identifikation von Onlineshops die Machine Learning Algorithmen „Neuronale Netze“, „Logistische Regression“ und „Random Forest“ verwendet und die Ergebnisse mit Maßzahlen für Präzision, Sensitivität und Richtigkeit überprüft. Dabei stellte sich heraus, dass die Logistische Regression als Algorithmus prädiktiver Modellierung keine größeren Fehlerquoten erzeugt hat als die bedeutend rechenaufwendigeren und komplizierten Alternativverfahren.

Trainingsregime

Ziel des Funktionslernens als Lernen ist es nun, mit h die möglichst gleichen Ergebnisse zu erzielen wie mit der Funktion f. Erreicht wird dies durch Anwenden der hypothetischen Funktion auf historische, elektronisch erhobene Unternehmensdaten mit bereits bekannten Unternehmenseigenschaften, die auf einen Trainings- und einen Testdatensatz aufgeteilt werden. Mit dem Verfahren des absteigenden Gradienten werden die Kausalitätsparameter durch Minimierung einer aus der hypothetischen Funktion abgeleiteten, konvexen, empirischen Fehlerfunktion iterativ mit den Trainingsdaten bestimmt/gelernt und über die Testdaten geprüft. Diese Art des Funktionslernens heißt Gradienten basiertes Lernen. Die Nutzung der Trainingsdaten beim vollüberwachten Funktionslernen erfolgt häufig im Batch-Modus. Das heißt, dass alle Datensätze der Trainingsdaten in einem Optimierungsprozess verwendet

werden, die vorher manuell recherchiert worden sind. Je mehr Fälle korrekt klassifiziert werden, umso besser ist die Funktion gelernt worden.

Bei dem hier vorgestellten Ansatz wird das iterative Online-Trainingsregime als teilüberwachtes Funktionslernen angewendet. Dies bedeutet, dass der Trainings- und der Testdatenbestand zwar initial ausschließlich durch manuelle Recherche entstehen, aber nicht statisch sind. Der Bestand wird vielmehr, abhängig von Prüfergebnissen des angewendeten prädiktiven Verfahrens, durch das automatisierte Hinzufügen von neuen Datensätzen maschinell ertüchtigt. Die Anzahl und Art der neuen Trainingsdatensätze folgt einer Funktion (siehe Bottou und Le Cun, 2004).

Nach dem Lernprozess werden im HSL-Verfahren Datensätze der Verknüpfungsdaten mit sehr hoher oder sehr niedriger prognostizierter Wahrscheinlichkeit für das Auftreten der Unternehmenseigenschaft den Trainingsdaten automatisiert hinzufügefügt und das prädiktive Verfahren erneut durchgeführt. Dadurch entsteht ein iterativer Lernprozess durch den die Kausalitätsparameter Prozedur für Prozedur angepasst werden.

Prognose

Mit den gelernten Kausalitätsparametern aus den Trainings- und Testverfahren und der hypothetischen Funktion können nun die Auftrittswahrscheinlichkeiten der interessierenden Eigenschaften für unbekannte Daten, nach elektronischer Extraktion der Erhebungsmerkmale bestimmt werden. Überschreiten die Wahrscheinlichkeiten einen vordefinierten Grenzwert, wird die interessierende Eigenschaft für die Webseite automatisch festgelegt (siehe Nilsson, 1998).

Anwendung auf Amtliche Daten

Technische Umsetzung

Im Hessischen Statistischen Landesamt ist von Oktober 2017 bis Mai 2018 eine für das Webscraping geeignete IT-Infrastruktur aufgebaut worden.

Dabei wurde ein vom italienischen, nationalen Statistikamt (ISTAT) bereitgestelltes Java-Programm zum Suchen und Extrahieren von Webseiten über Metasuchmaschinen und zur Speicherung und Trefferbewertung der enthaltenen Quelltexte, installiert, eingestellt und weiterentwickelt.

Für diesen Zweck ist auf einem virtuellen Server des Hessischen Statistischen Landesamts eine Datenbank mit den entsprechenden amtlichen Stammdaten errichtet worden. Zur Verknüpfung und Weiterverarbeitung der gefundenen Unternehmenswebseiten sowie für die Anreicherung mit Fachdaten, sind verschiedene R-Programme entwickelt worden.

Stammdaten aus dem amtlichen Unternehmensbestand

Zur Erstellung Amtlicher Statistiken werden Unternehmen in der Regel aus dem Bestand des statistischen Unternehmensregisters ausgewählt und nach entsprechenden fachthemenabhängigen, betriebsbezogenen Inhalten befragt.

Das statistische Unternehmensregister ist eine regelmäßig aktualisierte Datenbank mit Unternehmen und Betrieben aus nahezu allen Wirtschaftsbereichen mit steuerbarem Umsatz aus Lieferungen und Leistungen und/oder Beschäftigten. Die Quellen des Unternehmensregisters sind u. a. administrative Daten aus Verwaltungsbereichen wie der Bundesagentur für Arbeit oder von Finanzbehörden, und zum anderen Angaben aus einzelnen Bereichsstatistiken, beispielsweise aus Erhebungen des Produzierenden Gewerbes, des Handels oder des Dienstleistungsbereichs (siehe Statistisches Bundesamt, 2015).

Das statistische Unternehmensregister ist Grundlage fast jeder amtlichen Wirtschaftsstatistik und enthält wichtige Stammdaten wie Name, Adresse oder Steuernummer von etwa 340 000 in Hessen ansässigen Unternehmen.

Da die gegenwärtigen technischen Kapazitäten das Erheben und Verknüpfen von 100 Unternehmenswebseiten pro Tag erlaubten, war die automatisierte Erhebung des gesamten hessischen, statistischen Unternehmensregister nicht in einem angemessenen Zeitraum möglich. Als aktuelle Stichprobe des hessischen Unternehmensregisters waren jedoch die in der Erhebung zur Informations- und Kommunikationstechnologie in Unternehmen (IKT 2017) befragten 1658 Einheiten inhaltlich gut geeignet, da diese u. a. nach dem Vorhandensein einer unternehmenseigenen Webseite befragt wurden. Die Unternehmensmerkmale wurden in die Datenbank auf den Webserver geladen und das Webscraping-Verfahren auf die Datensätze angewendet.

Ergebnisse der Verknüpfungen

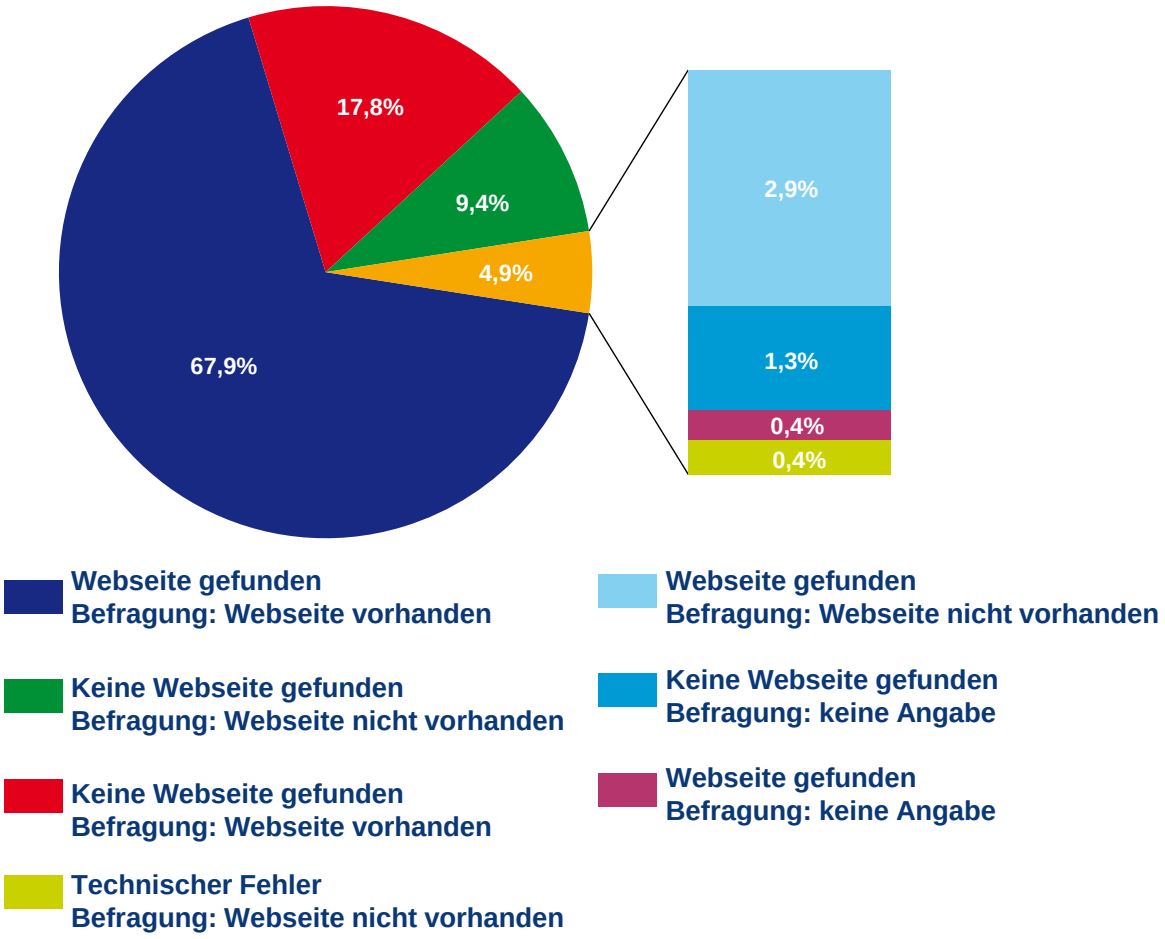
Abbildung 6 zeigt die Ergebnisse der Informationsextraktion durch Webscraping mit dem Bestand der 1658 Unternehmen aus dem Unternehmensregister, die für die IKT 2017 nach Informations- und Kommunikationstechnologie, u. a. nach dem Betrieb einer unternehmenseigenen Webpräzens, befragt worden sind. Inhaltlicher Gegenstand der Illustration ist, ob eine Webpräsenz zugeordnet worden ist und ob das jeweilige Unternehmen eine unternehmenseigene Webseite in der Befragung IKT 2017 angegeben hat.

Bei 77,3 % der befragten Unternehmen stimmte das Ergebnis der Befragung mit den Ergebnissen des Webscraping überein. Darunter waren 67,9 % mit Webpräzens und 9,4 % ohne. Von 1420 Unternehmen, die angegeben haben, eine Webseite zu betreiben wurde für 79,2 % eine Zuordnung erzielt.

Für 21,1 % der Unternehmen wichen die Webscraping-Zuordnungen von den Befragungsergebnissen ab. Darunter waren 18,2 % (17,8 % nicht erkannt zuzüglich 0,4 % technischer Fehler) nicht erkannte Webpräsenzen und 2,9 % erzielte Zuordnungen, obwohl in der Befragung angegeben wurde, dass keine Webpräsenz vorhanden ist. Etwa 1,7 % der Unternehmen machten keine Angabe. Von 211 Unternehmen, die angegeben haben, über keine Webseite zu verfügen, wurde für 22,7 % durch das Webscraping trotzdem eine Zuordnung erzielt.

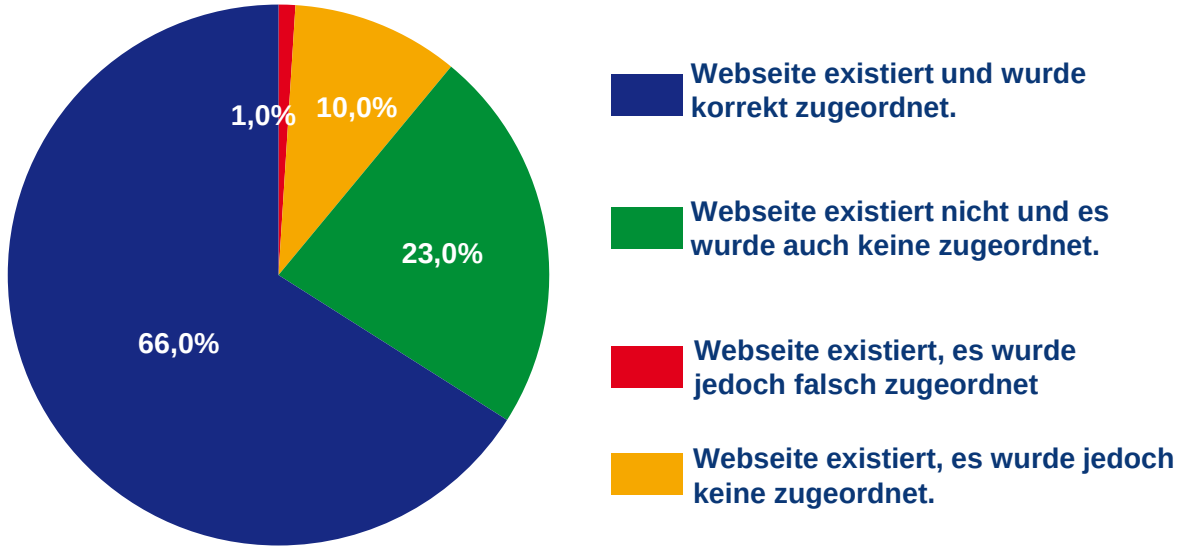
In 48 Fällen haben Unternehmen angegeben, über keine Webpräsenz zu verfügen, obwohl diese nach Überprüfung vorhanden war und durch das Webscraping richtig zugeordnet wurde. Als Grund für die Abweichungen zwischen Befragungsergebnis und Webscraping können die unterschiedlichen Erhebungszeitpunkte nicht ausgeschlossen werden.

Abbildung 6: Ergebnisse der Zuordnung von extrahierten Webseiten zu 1658 hessischen Unternehmen aus dem amtlichen Datenbestand



Insgesamt sind für 71,5 % der Unternehmen somit 1186 Webpräsenzen automatisiert zugeordnet worden. Unabhängig davon, ob Unternehmen angegeben haben, eine Webseite zu betreiben oder nicht, muss beurteilt werden, ob die Zuordnungen korrekt waren. Daher wurde für eine Zufallsstichprobe von 100 Unternehmen die Zuordnung durch manuelle Recherche überprüft. Die Abbildung 7 illustriert die Ergebnisse dieser Überprüfung.

Abbildung 7: Ergebnisse der Überprüfung von 100 Zuweisungen extrahierter Unternehmenswebseiten



In 89,0 % der überprüften Fälle wurden korrekte Zuweisungen erzielt (Richtigkeitsrate). Darunter waren 66,0 % erkannte Onlinepräsenzen und 23,0 % Unternehmen ohne eigene Webseite. Die recherchierten Webpräsenzen von Unternehmen mit eigener Webseite, wurden zu 85,7 % richtig erkannt (Sensitivitätsrate). Die automatisierten positiven Zuordnungen von Unternehmenswebseiten haben sich dabei zu 98,5 % (Präzisionsrate) als korrekt herausgestellt.

Bei 11,0 % der Fälle hat das Verfahren zu falschen Zuordnungen geführt (Fehlklassifikationsrate). Darunter wurde bei 1,0 % zwar das Vorhandensein einer unternehmenseigenen Webseite erkannt, jedoch die falsche Onlinepräsenz ermittelt. Bei 10,0 % der Fälle, in denen sich keine Webpräsenzen recherchieren ließen, wurden Zuteilungen fehlerhaft erzielt.

Die verschiedenen beschriebenen Prüfungsergebnisse der binären Klassifikation machen es für die Interpretation der Resultate nötig, ein Maß zur Beurteilung der Zuordnungsgüte als relative Korrektheitsintensität berechnen zu können. Hier bietet sich der F-Wert, errechnet als das harmonische Mittel aus Sensitivitätsrate und Präzisionsrate an.

Der F-Wert beträgt bei der niedrigsten Korrektheit 0 und bei höchster Richtigkeit 1. Bei der Anwendung im HSL lag er bei 0,917, was für eine sehr hohe Korrektheit der durchgeführten Klassifikation spricht.

Von den 1186 identifizierten und verknüpften Webseiten von hessischen Unternehmen der IKT 2017 konnten etwa 1111 heruntergeladen und die Quelltexte analysiert werden. Dabei wurden Schlüsselwörter aus dem Bereich „Handel“ in verschiedenen Gruppen zusammengefasst und deren Vorkommen auf den heruntergeladenen Quelltexten geprüft. Das Vorkommen von Wörtern jeweiliger Wortgruppen wurde innerhalb von verschiedenen Indikatorvariablen festgehalten.

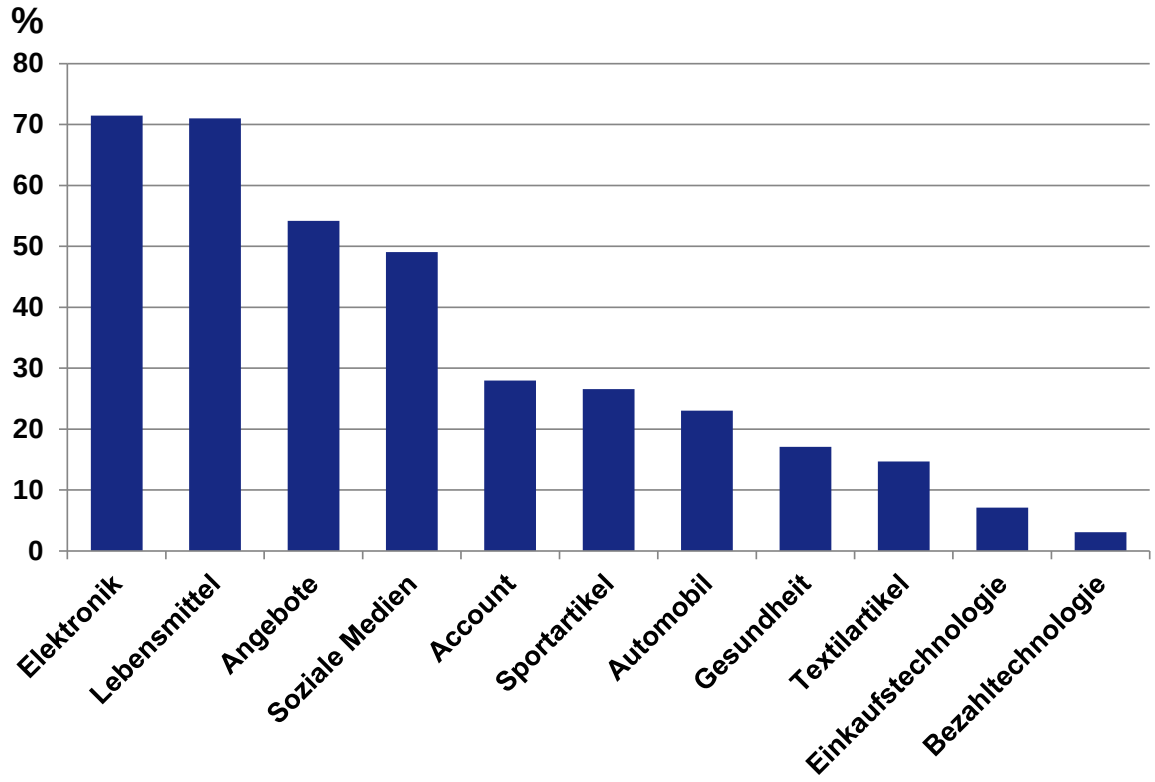
Die Resultate sind in Abbildung 8 dargestellt und zeigen, dass die meisten Schlüsselwörter im Bereich „Elektronik“ (71,5 %) und „Lebensmittel“ (71,0 %), gefolgt von Wörtern aus dem

Bereich „Sonderangebot“ (54,2 %) gefunden worden sind. Etwa auf der Hälfte (49,1 %) aller analysierten Webseiten wurden Bezüge zu sozialen Medien gefunden.

Weitere Schlüsselwortgruppenvorkommen mit Häufigkeiten über 10 % lagen im Bereich „Account-Technologie (28,0 %)“, „Sportartikel“ (27,0 %), „Automobil“ (23,0 %), „Gesundheit“ (17,1 %) und „Textilartikel“ (14,7 %).

Die stark mit der Eigenschaft eines Onlineshops im Zusammenhang stehenden Schlüsselwortgruppen der Bezahltechnologie („Visa“, „Kreditkarte“, „PayPal“, „EC-Karte“ und Ähnliches) und der Einkaufstechnologie („Warenkorb“, „Einkaufswagen“, „Shopping Cart“ und Ähnliches) waren bei unter 10 % aller zugeordneten Unternehmenswebseiten zu finden.

Abbildung 8: Auftreten ausgewählter Schlüsselwortgruppen auf n = 1111 auslesbaren Webseiten von hessischen Unternehmen der IKT 2017



Maschinelle Bestimmung der latenten Unternehmenseigenschaft „E-Commerce“

Es gibt Unternehmenseigenschaften, wie die Beteiligung an E-Commerce-Aktivitäten, die für die amtliche Statistik von großem Interesse sind. Diese Eigenschaft ist durch das Vorhandensein eines Onlineshops erfüllt. Oft zeigt der erste Blick auf eine Webseite mit einem aktiven Onlineshop, über den bestellt und bezahlt werden kann, dass sich die Eigenschaft nicht sofort von der Webseite ablesen lässt. Der Betrieb eines Onlineshops für den elektronischen Versandhandel ist somit nicht unbedingt direkt und zweifelsfrei aus den Quelltexten der zugeordneten Unternehmenswebseiten elektronisch auslesbar. Es gibt jedoch Merkmale, u. a. das Vorhandensein von Online-Zahlungs-Optionen oder eines Warenkorbs, die mit dem Betrieb eines Onlineshops in einer schätzbaren Verbindung stehen können.

Für die Durchführung der maschinellen Ermittlung der latenten Eigenschaft „Onlineshop“ mussten zunächst geeignete Prädiktoren zur Stützung der erwähnten Zusammenhangshypothese unter den automatisiert erhobenen Merkmalen gefunden werden. Darüber hinaus war die Verfügbarkeit von historischen Daten, in denen die Eigenschaft „Onlineshop“ bereits bekannt war, erforderlich. Für diesen Zweck war eine Stichprobe von zunächst 146 Handelsunternehmen aus den erhobenen Daten gezogen worden. Anschließend ist mit Hilfe manueller Recherche ein binärer Klassifizierer generiert worden, der einen positiven Fall bei beobachten eines Webshops enthielt und einen negativen Fall, wenn sich der jeweiligen Webseite kein Webshop entnehmen ließ.

Eine Assoziationsanalyse mit den historischen Daten ergab signifikante, mittelstarke Zusammenhänge zwischen dem Vorhandensein eines Onlineshops auf den verknüpften Unternehmenswebseiten und den folgenden erhobenen Merkmalen:

- Vorhandensein einer Einkaufstechnologie (z. B. Shopping Cart, Warenkorb, Login),
- Vorhandensein einer Bezahltechnologie (z. B. Visakarte, PayPal, Sofortüberweisung).

Mit den geeigneten Prädiktoren wurde nun ein prädiktives, auf Wahrscheinlichkeiten basierendes Modell, wie in den vorherigen Punkten beschrieben, als Verfahren des maschinellen Lernens zum automatisierten Ermitteln von Onlineshops gewählt. Hierfür wurden die historischen Daten zu jeweils 50 % in eine Trainings- und eine Teststichprobe zufällig aufgeteilt.

Mit dem prädiktiven Modell wurden nun die Kausalitäten zwischen den Prädiktoren und dem Klassifizierer für die Eigenschaft „Webshop“ in der Trainingsstichprobe in einem ersten Lernprozess ermittelt und über die Teststichprobe geprüft. Anschließend, wurden mit Hilfe der gelernten Kausalitäten aus dem Datenbestand der verknüpften Unternehmenswebseiten 15 Datensätze mit einer geschätzten Wahrscheinlichkeit für die Eigenschaft „Onlineshop“ von mindestens 70 % ausgewählt und der historischen Datenstichprobe hinzugefügt. Danach wurde der gesamte Lernprozess mit einem neuen Datenbestand aus Trainings- und Teststichprobe von nun 161 Beobachtungen für einen weiteren Lernprozess noch einmal durchgeführt[1].

Mit den geschätzten Kausalitäten nach zwei Lernprozessen war es möglich, eine Klassifikation für alle zugeordneten 1111 Webseiten von Unternehmen der IKT 2017 nach der Eigenschaft „Onlineshop“ durchzuführen. Dabei wurde für etwa 5 % der Unternehmenswebseiten ein Onlineshop vorhergesagt.

Für die nach zwei Lernprozessen überprüften 161 Zuweisungen von Onlineshops ließen sich die Ergebnisse nun wiederrum überprüfen. Abbildung 9 und 10 zeigen die Ergebnisse dieser Prüfung.

Beim Lernprozess 1 wurden in 87,7 % der überprüften Fälle korrekte Prognosen erzielt (Richtigkeitsrate). Darunter waren 5,5 % erkannte Onlineshops und 82,2 % Unternehmenswebseiten ohne Onlineshop. Die recherchierten Onlineshops wurden nur zu 30,8 % richtig

[1] Die Anzahl der Lernprozesse ist nicht beschränkt. Um für diese Studie alle automatisch zugeordneten Webshops auch manuell überprüfen zu können, wurde die Anzahl der Lernprozesse auf 2 beschränkt.

erkannt (Sensitivitätsrate). Der Anteil korrekter Zuweisungen von Onlineshops an allen positiven Klassifizierungen betrug jedoch 100,0 % (Präzisionsrate).

Bei 12,3 % der Fälle führte das Verfahren zu falschen Prognosen und hat existierende Onlineshops nicht erkannt (Fehlklassifikationsrate).

Der F-Wert beträgt hier 0,471, was für eine deutlich geringere Korrektheit der durchgeführten Identifikation von Webshops spricht als bei der Zuweisung von Unternehmenswebseiten.

Abbildung 9: Ergebnisse der Überprüfung maschinell zugeordneter Unternehmenseigenschaften: „Onlineshop“

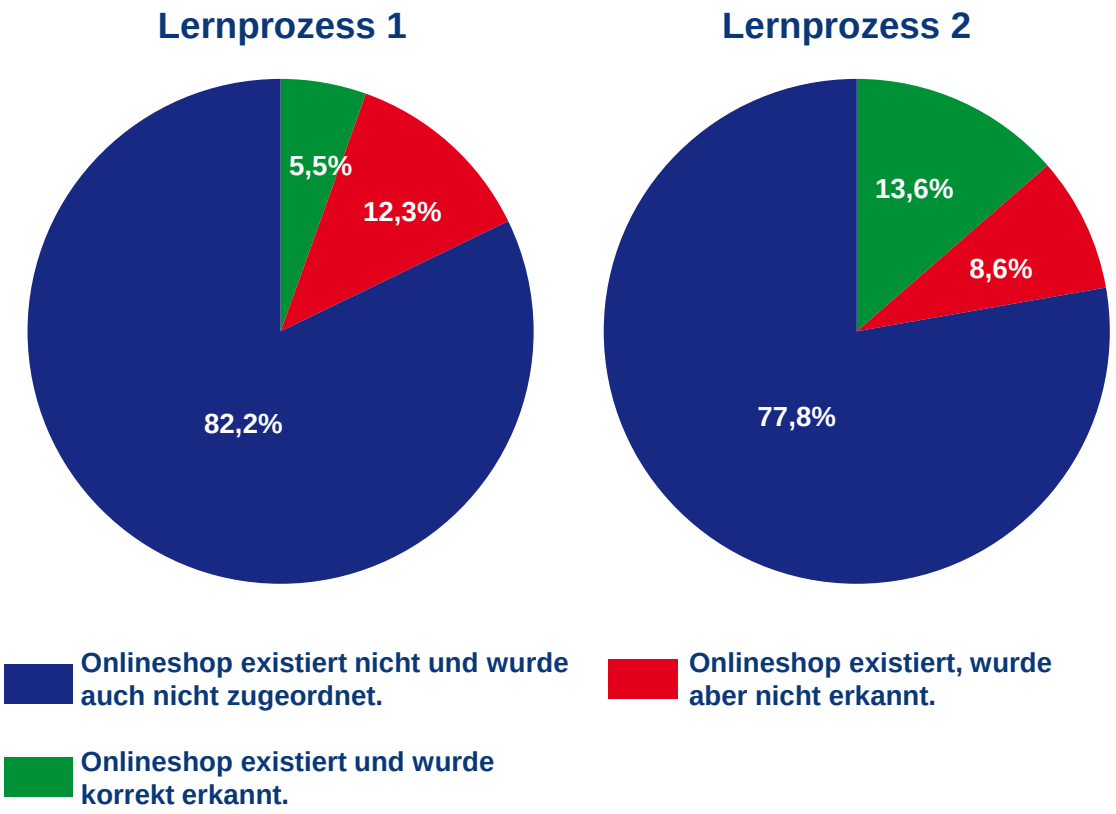
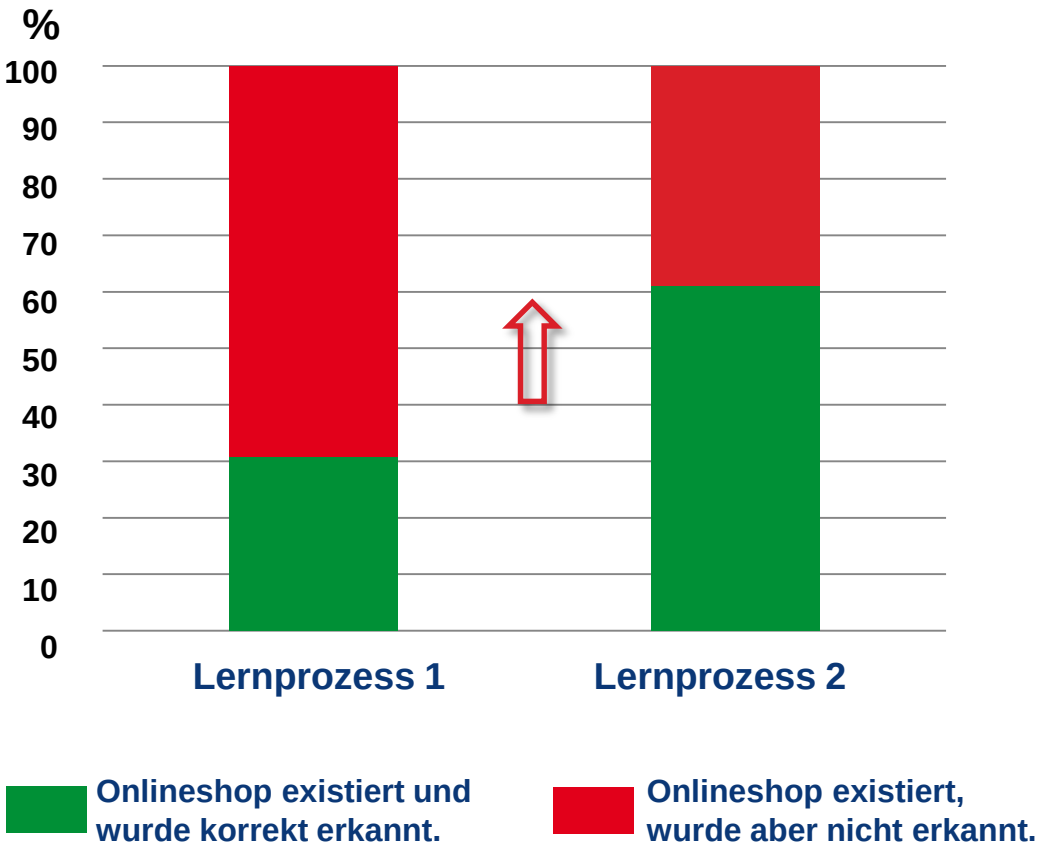


Abbildung 10: Anteil korrekter Klassifizierungen von Onlineshops



Beim Lernprozess 2 wurden in 91,4 % (+ 3,7 %) der überprüften Fälle korrekte Prognosen erzielt (Richtigkeitsrate). Darunter waren nun 13,6 % erkannte Onlineshops und 77,8 % Unternehmenswebseiten ohne Onlineshop.

Die recherchierten Onlineshops, wurden jetzt zu 61,1 % richtig erkannt (Sensitivitätsrate). Somit sind durch den zweiten Lernprozess die Sensitivitätsrate und damit die Erkennung von Webshops um 30,3 % gestiegen. Der Anteil korrekter Zuweisungen von Onlineshops an allen Klassifizierungen betrug wiederum 100,0 % (Präzisionsrate).

Die Fehlklassifikationsrate ist um 3,7 % auf 8,6 % gesunken.

Der F-Wert ist im zweiten Lernprozess um 0,288 Einheiten auf 0,759 gestiegen was für ein deutlich besseres Ergebnis im Vergleich zum ersten Lernprozess spricht.

Das Verfahren zeigt, dass der Lernprozess von den Ergebnissen im Zieldatensatz profitieren sollte, da dieser so treffsicherer wird. Mit fortlaufenden Iterationen sollte die Anzahl der hinzufügbaren neuen Datensätze aus dem Zieldatensatz zu einem gegeben Stand sinken. Das Verfahren würde somit stoppen, wenn für keinen weiteren Datensatz eine Wahrscheinlichkeit prognostiziert werden könnte, welche die vordefinierten Grenzwerte überschreitet bzw. unterschreitet.

Ermitteln von Schlafgelegenheiten von Beherbergungsbetrieben des HRS-Portals

Für ein Methodenprojekt wurde die Bettenanzahl von Beherbergungsbetrieben im Raum München benötigt und als Quelle für die Unternehmensidentifikation das im Internet öffentlich zugängliche Portal „hrs.de“ gewählt.

Auf dem HRS-Portal sind folgende auslesbare und frei zugängliche Informationen abrufbar:

Anzahl Einzelzimmer, Anzahl Doppelzimmer, Anzahl Zimmer allgemein, Name, Anschrift (Straße, Hausnummer, Postleitzahl), Geokoordinaten (Längengrad, Breitengrad), Hotelkategorie (Anzahl Sterne), Hotelart und das Rating.

Die Struktur der jeweiligen Sekundärwebseiten mit den Informationen über den entsprechenden Beherbergungsbetrieb folgt einem ähnlichen Muster wie die unter der Google-Suchmaschine enthaltenen Webseiten.

Ähnlich dem Verfahren, welches den Google-Resultaten folgt, werden hier die Quelltexte hierarchisch analysiert, bis die Ergebnisquelltexte/Hotelseitenquelltexte gefunden wurden. In diesen wird kontextabhängig mit Methoden des Text-Minings nach den Ausprägungen der jeweiligen Merkmale gesucht, um diese strukturiert zu speichern.

Auf diese Weise konnten alle Münchener Beherbergungsbetriebe am Stichtag 6. Juli 2018, die über das HRS-Portal gelistet wurden, ausgelesen und die enthaltenen Daten über das jeweilige Bettenangebot gespeichert werden. Der Datensatz enthält statistisch quantifizierbare Informationen (Ausprägungen) aller oben beschriebenen Merkmale.

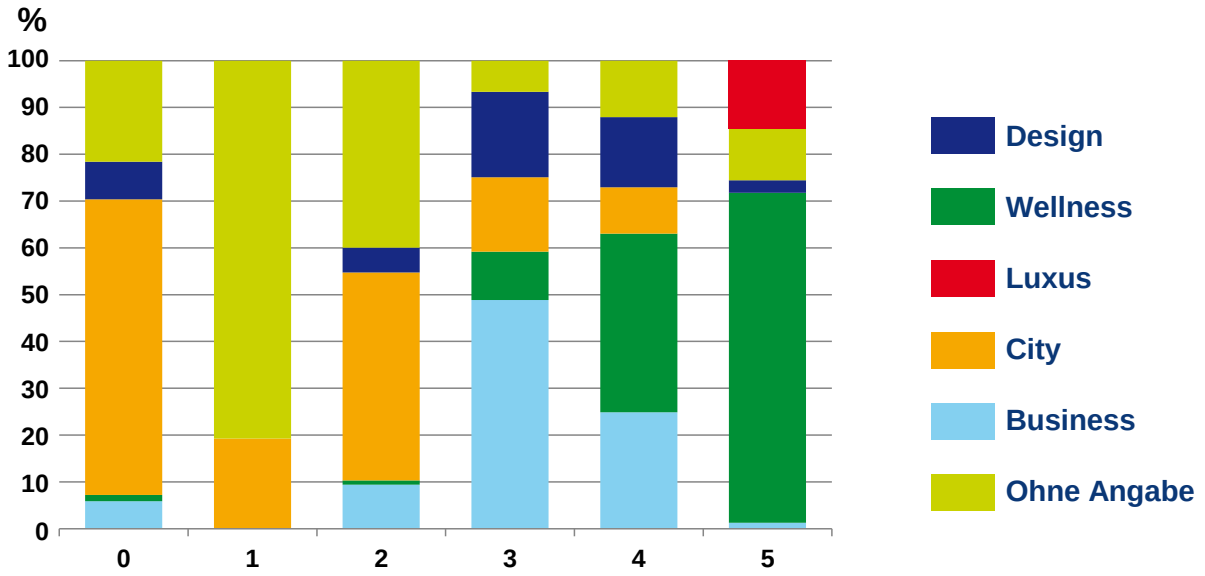
Im HRS-Portal wurden am 6. Juli 2018 37 562 Betten in München angeboten. 54,0 % davon entfielen auf Doppelzimmer. 25,0 % des Bettenangebots entfiel auf Einzelzimmer und 21,0 % auf andere nicht definierte Angebotsvarianten.

Darüber hinaus waren noch weitere statistisch auswertbare Mikrodateninformationen auf dem Onlineportal enthalten. So waren u. a. die Hotelkategorie, und der Hoteltyp verfügbar. Die Abbildung 10 zeigt daher die Verteilung der 37 562 Schlafgelegenheiten im Raum München auf die erwähnten Merkmale.

Die Ergebnisse zeigen, dass Wellness und City-Hotels in München am meisten vertreten sind. Dabei sinkt die relative Bedeutung der City-Hotels mit steigender Hotelkategorie (Bewertung nach Sternen). Bei Wellness-Hotels ist der Zusammenhang umgekehrt. Diese sind erst ab drei Sternen beobachtbar und nehmen in der Auftrittshäufigkeit von 70,5 % den größten Anteil der Münchener 5 Sternehotels im HRS-Portal ein.

Luxushotels sind mit ca. 14,6 % Anteil an den Schlafgelegenheiten ausschließlich im obersten Sternebereich zu finden. Business Hotels dominieren mit 48,9 % der angebotenen Schlafgelegenheiten in der mittleren Hotelkategorie mit 3 Sternen. Hotels ohne Angabe bzgl. Hoteltyp sind am meisten in der Hotelkategorie 1 Stern mit etwa 80,8 % Anteil zu finden.

Abbildung 11: Verteilung der Schlafgelegenheiten des HRS-Portals im Raum München auf Hotelkategorie in Sternen und Hoteltyp



Die Eigenschaft City-Hotel ist mit etwa 63,2 % am häufigsten bei Hotels beobachtbar, die nicht am internationalen Verfahren zur Hotelkategorisierung teilnehmen.

Das Webscraping von Online-Portalen zeigt, dass es sehr gut geeignet ist, umfassende Informationen über Unternehmen einer bestimmten Branche elektronisch zu erheben, ohne die amtlichen Register dafür benutzen zu müssen.

Es zeigt auch, dass das Ermitteln von Unternehmenseigenschaften, wie die in diesem Fall interessierende Anzahl der Betten in Abhängigkeit von der Verfügbarkeit, relativ einfach und treffsicher elektronisch erhoben werden könnten.

Fazit und Verbesserungspotenzial

Die Ergebnisse des ersten Versuchs (Webscraping + prädiktive Modellierung) verdeutlichen, dass das Potenzial zum Zuordnen von Unternehmenswebseiten bezogen auf die Treffsicherheit sehr gut funktioniert hat. Die Ergebnisse sprechen zum größten Teil für ein bedeutendes Potenzial des Webscraping.

Bei der prädiktiven Modellierung hängt die Treffsicherheit im Wesentlichen von dem Umfang und der Zusammensetzung des Trainingsdatensatzes und von der Auswahl der Schlüsselwörter ab. Hier müsste noch viel konzeptionelle Vorarbeit geleistet werden, um qualitativ hochwertige Trainingsdatensätze erzeugen zu können. Das Verfahren hat gezeigt, dass ein dynamisch gestalteter, teils automatisiert ertüchtigter Trainingsdatensatz die Treffsicherheit des teilüberwachten Lernverfahrens deutlich erhöhen kann. Das im HSL über zwei Lernprozesse durchgeführte Verfahren der prädiktiven Modellierung wird zukünftig für eine unbeschränkte Anzahl an Iterationen weiterentwickelt, um eine optimale Treffsicherheit bei der binären Klassifikation zu erzielen.

Die Art der Schlüsselwörter, die Größe des Trainingsdatenbestands und die in Frage kommenden Prädiktoren für das maschinelle Lernen sind stark abhängig von der zu prognostizie-

renden Unternehmenseigenschaft. Ein manuell recherchierter und mit bestimmten erhobenen Merkmalen zusammengeführter Trainingsdatenbestand für das Thema „E-Commerce“ kann nicht verwendet werden, um bspw. Unternehmen mit grüner Technologie oder Freiberufler in der Kreativwirtschaft automatisiert zu ermitteln.

Um Unternehmenseigenschaften mit Methoden des maschinellen Lernens zu ermitteln, wird es je nach Themeninteressen und fachstatistischen Fragestellungen bis auf weiteres notwendig sein, Trainingsdatensätze mit Hilfe manueller Recherche zu erstellen, zu validieren, und diese zu pflegen. Die im HSL genutzte Online-Methode zum automatisierten Anreichern der Trainingsdaten war hierbei sehr hilfreich und birgt, u. a. im Hinblick auf das Einbringen von Modelldiagnostik für eine maschinelle Prädiktorenauswahl, noch einiges an Entwicklungspotential.

Die technische und methodologische Umsetzung eines Verfahrens des maschinellen Lernens zum prädiktiven Identifizieren von latenten Unternehmenseigenschaften hat gut funktioniert. Hier ist die Verwendung einer R-Umgebung sehr komfortabel und ausreichend schnell.

Das Durchsuchen und Speichern von Unternehmenswebseiten von Onlineportalen ist aufgrund der gleichbleibenden Struktur der Webseiten innerhalb des jeweiligen Portals recht einfach und treffsicher durchführbar. Kommerzielle Onlineportale müssen zwar manuell recherchiert werden, je nach Wirtschaftszweig können diese jedoch in einer Datenbank hinterlegt und bei Bedarf abgerufen werden. Das Scraping von Onlineportalen ohne das Nutzen einer vorgeschalteten Metasuchmaschine hat sich als sehr schnell herausgestellt. Das Ermitteln aller über das HRS-Portal gelisteten Münchener Beherbergungsbetriebe hat mit dem geschriebenen R-Algorithmus etwa 5 Minuten gedauert. Schon aufgrund der Unabhängigkeit von Suchmaschinen wie Google.com ergibt sich hier zunächst kein kapazitätsspezifisches, großes Verbesserungspotenzial. Weitere Untersuchungen werden zeigen inwieweit innerhalb eines Gesamtalgorithmus zunächst bereits bekannte Onlineportale abgesucht, verknüpft und erst im Anschluss daran das Webscraping mit Metasuchmaschine auf die noch nicht verknüpften Unternehmensdatensätze angewendet werden könnten.

Für das Webscraping von Unternehmenswebseiten sieht dies jedoch anders aus. Bei den derzeit gegebenen Kapazitätsgrenzen würde die Vollerhebung unternehmensbezogener Webseiten der etwa 300 000 hessischen Unternehmen der Unternehmensregisterkopie 8,3 Jahre dauern. Die Weiterentwicklung des Webscraping wird daher stark auf die Steigerung der Kapazitäten ausgerichtet sein, um die vollständige Erhebung des hessischen Unternehmensregisters durch das Webscraping innerhalb von etwa 30 Tage durchführen zu können. Dies wäre ein zufriedenstellender Zeitraum.

Literaturverzeichnis

BARCAROLI, Giulio, Monica SCANNAPIECO und 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), S. 25-41. [Zugriff am: 03.09.2018]. RePEc. ISSN: 0034-6535, Verfügbar unter: http://www.sieds.it/listing/RePEc/journl/2016LXX_N4_RIEDS_25-41_Scannapieco.pdf

BOTTOU, Leon, und 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), S. 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), S. 62-512. [Zugriff am: 01.09.2018]. Science Direct. ISSN: 0304-3932. Verfügbar unter: DOI: 10.1016/j.jmoneco.2012.10.002

COHEN, William W., Pradeep RAVIKUMAR und 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., und 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. [Zugriff am: 01.09.2018]. Sciedu. ISSN: 1927-6982. Verfügbar unter: DOI: 10.5430/air.v2n1p44

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

FREES, Edward W., Richard A. DERRIG und Glenn MEYERS, 2014. *Predictive Modeling Applications in the Actuarial Science – Volume I: Predictive Modeling Techniques*. New York: Camebridge 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 [Zugriff am: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Verfügbar unter: DOI: 10.3233/SJI-160965

HOEKSTRA, Rutger, Olav TEN BOSCH und Frank HARTEVELD, 2012. Automated data collection from web sources for official statistics: First experiences. In: *Statistical Journal of the IAOS* [online]. **28**(3,4), S. 99-111 [Zugriff am: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Verfügbar unter: 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, und 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. [Zugriff am 04.09.2018]. Verfügbar unter: http://robotics.stanford.edu/people/nilsson/mlbook.html

OOSTROM, Lotte, Adam N. WALKER, Bart STAATS, Magda SLOOTBEEK-VAN LAAR, Shirley O. AZURDUY und 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 und 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), S. 165-176 [Zugriff am: 02.09.2018]. IOS Press Content Library. ISSN 1875-9254. Verfügbar unter: 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, S. 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, S. 22.

STATISTISCHES BUNDESAMT, 2015. *Unternehmen und Betriebe im Unternehmensregister: Methodische Grundlagen, Definitionen und Qualität des statistischen Unternehmensregisters*. Statistisches Bundesamt. [Zugriff am: 05.09.2018]. Verfügbar unter: 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, und Cosimo PALMISANO, 2008. Using Context to Improve Predictive Modeling of Customers in Personalization Applications. In: *IEEE Transactions on Knowledge and Data Engineering* [online], **20**(11), S. 1535-1549. [Zugriff am: 05.09.2018]. ISSN 1041-4347. Verfügbar unter: http://doi.ieeecomputersociety.org/10.1109/TKDE.2008.110

US AIR FORCE, 2006. *Method and Apparatus for improved Web Scraping*. Erfinder: SALERNO, John und Douglas M. BOULWARE. 04.07.2006. Anmeldung: 26.08.2004. US, Patentschrift US7072890B2

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

Impressum

Herausgeber:
Hessisches Statistisches Landesamt,
Wiesbaden, Rheinstraße 35/37
Telefon: 0611 3802-0,
Telefax: 0611 3802-890
E-Mail: info@statistik.hessen.de
Internet: https://statistik.hessen.de

Ansprechpartner:
Normen Peters
Telefon: 0611 3802-517
E-Mail: normen.peters@statistik.hessen.de