Search Studies an der HAW Hamburg

Dirk Lewandowski*, Alexandra Krewinkel, Mareike Gleissner, Dorle Osterode, Boris Tolg, Martin Holle und Sebastian Sünkler

Entwicklung und Anwendung einer Software zur automatisierten Kontrolle des Lebensmittelmarktes im Internet mit informationswissenschaftlichen Methoden

https://doi.org/10.1515/iwp-2019-0005

Zusammenfassung: In diesem Artikel präsentieren wir die Durchführung und die Ergebnisse eines interdisziplinären Forschungsprojekts zum Thema automatisierte Lebensmittelkontrolle im Web. Es wurden Kompetenzen aus den Disziplinen Lebensmittelwissenschaft, Rechtswissenschaft, Informationswissenschaft und Informatik dazu genutzt, ein detailliertes Konzept und einen Software-Prototypen zu entwickeln, um das Internet nach Produktangeboten zu durchsuchen, die gegen das Lebensmittelrecht verstoßen. Dabei wird deutlich, wie ein solcher Anwendungsfall von den Methoden der Information-Retrieval-Evaluierung profitiert, und wie sich mit relativ geringem Aufwand eine flexible Software programmieren lässt, die auch für eine Vielzahl anderer Fragestellungen einsetzbar ist. Die Ergebnisse des Projekts zeigen, wie komplexe Arbeitsprozesse einer Behörde mit Hilfe der Methoden von Retrieval-Tests und gängigen Verfahren aus dem maschinellen Lernen effektiv und effizient unterstützt werden können.

Deskriptoren: Lebensmittelwissenschaft, Lebensmittelüberwachung, Kontrolle, Informationswissenschaft, Internet, Recherche, Softwareentwicklung, Interdisziplinär, Webrecherche, Retrieval, Maschinelles Lernen

Development and application of software for automated control of the food market on the Internet using information science methods

Abstract: In this paper we present the implementation and results of an interdisciplinary research project on the topic of using automation to support food control on the web. Competences from the disciplines of food science, law, information science and computer science were used to develop a detailed concept and a software prototype to search the internet for product offers that violate food law. We show how such an application benefits from the methods of information retrieval evaluations and how a flexible software can be programmed with little effort by using common patterns in software development and open source libraries. We also show how the software can be used to carry out other projects. The results of our research demonstrate how complex tasks can be supported effectively and efficiently by using methods of retrieval tests and common machine learning processes.

Descriptors: Food Science, Law, Control, Information Science, Internet, Research, Software Development, Interdisciplinary, Web Research, Retrieval, Machine Learning

***Kontaktperson: Prof. Dr. Dirk Lewandowski,** Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information, Finkenau 35, 22081 Hamburg,
E-Mail: dirk.lewandowski@haw-hamburg.de
Alexandra Krewinkel, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Department Ökotrophologie, Ulmenliet 20, 21033 Hamburg,
E-Mail: alexandra.krewinkel@haw-hamburg.de
Mareike Gleissner, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Department Ökotrophologie, Ulmenliet 20, 21033 Hamburg,
E-Mail: mareike.gleissner@haw-hamburg.de
Dorle Osterode, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Department Ökotrophologie, Ulmenliet 20, 21033 Hamburg,
E-Mail: dorle.osterode@haw-hamburg.de
Prof. Dr.-Ing. Boris Tolg, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Department Ökotrophologie, Ulmenliet 20, 21033 Hamburg, E-Mail: boris.tolg@haw-hamburg.de
Prof. Dr. Martin Holle, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Department Ökotrophologie, Ulmenliet 20, 21033 Hamburg, E-Mail: martin.holle@haw-hamburg.de
Sebastian Sünkler, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information, Finkenau 35, 22081 Hamburg, E-Mail: sebastian.suenkler@haw-hamburg.de

Développement et application d'un logiciel de contrôle automatisé du marché alimentaire sur Internet à l'aide de méthodes de science de l'information

Résumé : Dans cet article, nous présentons la mise en œuvre et les résultats d'un projet de recherche interdisciplinaire sur le contrôle automatisé des aliments sur le Web. Les compétences des disciplines des sciences alimentaire, de jurisprudence, des sciences de l'information et de l'informatique ont été utilisées pour développer un concept détaillé et un prototype de logiciel permettant de rechercher sur Internet des offres de produits qui violent la législation alimentaire. Il devient clair comment un tel cas d'application bénéficie des méthodes d'évaluation de la recherche d'information, et comment un logiciel flexible peut être programmé avec relativement peu d'effort, qui peut également être utilisé pour de nombreuses autres tâches convenable. Les résultats du projet montrent comment les processus de travail complexes d'une autorité publique peuvent être soutenus de manière efficace et efficiente à l'aide de méthodes communes de tests de récupération et de procédures communes d'apprentissage automatique.

Descripteurs : Science alimentaire, Jurisprudence, Contrôle, Science de l'information, Internet, Recherche, Développement de logiciels, Interdisciplinaire, Recherche sur le Web, Recherche, Apprentissage machine pédagogique

Einleitung

Der Handel mit Waren über das Internet ist ein wachsender Markt in Deutschland (bevh, 2018). Längst sind Lebensmittel nicht länger Nischenprodukte. Dies liegt nicht zuletzt am Markteinstieg großer bekannter Anbieter wie Amazonfresh, Hellofresh.de und REWE online, die mit einem Umsatz von 50–200 Mill. Euro einen nicht unerheblichen Anteil ihrer Umsätze im Onlinehandel mit Lebensmitteln erzielen (ebd.). Zurzeit kaufen bereits ca. drei Millionen Deutsche online Lebensmittel ein (IfD, 2018).

Diese Entwicklungen werden auch von den Lebensmittelüberwachungsbehörden seit einigen Jahren mit erhöhter Aufmerksamkeit verfolgt. Neben Gesetzesänderungen, die spezielle Regelungen für den Onlinehandel beinhalten, gibt es auf europäischer Ebene neu eingerichtete Arbeitsgruppen zum Thema E-Commerce. Diese sollen Best-Practice-Strategien für die Kontrolle des Onlinehandels mit Lebensmitteln entwickeln.

In Deutschland hat sich dieses Themas die gemeinsame Zentralstelle „Kontrolle der im Internet gehandelten Erzeugnisse des LFGB[1] und Tabakerzeugnisse“ (G@ZIELT) angenommen, eingerichtet vom Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL). G@ZIELT wurde im Jahr 2016 als Zentralstelle der Bundesländer zur Kontrolle der im Internet gehandelten Erzeugnisse gegründet (BVL, o. J.). Zu den Hauptaufgaben gehören die Recherche nach nicht verkehrsfähigen Onlineangeboten und die Übermittlung der Händlerinformationen an die zuständigen Vor-Ort-Behörden. Diese Recherchen werden zurzeit von Mitarbeiterinnen und Mitarbeitern der Zentralstelle von Hand unter Zuhilfenahme von Suchmaschinen durchgeführt.

In diesem Aufsatz beschreiben wir die Konzeption und Entwicklung eines Systems, das Lebensmittelkontrolleure bei ihrer Arbeit unterstützt, indem verdächtige Händler im Internet automatisch identifiziert und zur manuellen Kontrolle vorgeschlagen werden. Das System greift dabei auf Daten aus kommerziellen Suchmaschinen zurück, so dass der Aufwand für ein eigenes Web-Crawling entfällt. Wir stellen zuerst den Hintergrund der Lebensmittelkontrolle im Internet aus der Sicht der Lebensmittelüberwachung dar und gehen dann auf die Konzeption und Umsetzung einer Software zur Unterstützung der Lebensmittelüberwachung ein. Im Fazit beleuchten wir, warum dieser spezielle Anwendungsfall auch ein Thema für die Informationswissenschaft ist und heben hervor, dass die Entwicklung einer Software zur Untersuchung solcher Fragestellungen mit vertretbarem Aufwand realisierbar ist.

Das Projekt AAPVL

Die Kontrolle des Lebensmittelmarktes im Web stellt die Überwachungsbehörden vor große Herausforderungen, die sich u. a. durch die immer weiter wachsende Menge an Daten und die Schnelllebigkeit des Internets ergeben. Dies war der Grund für die Durchführung des Forschungsprojekts „Entwicklung von automatisierten Analyseverfahren zur Identifizierung und Bewertung von nicht verkehrsfähigen Produkten des virtuellen Lebensmittelmarktes“ (AAPVL), in dem ein IT-basiertes System zur Unterstützung der Online-Lebensmittelkontrolle bei den Recherchetätigkeiten und der Datenverarbeitung ausgearbeitet werden sollte. Im Rahmen des Projekts wurden dafür ein Konzept zur Kontrolle des Online-Handels mit Lebensmitteln und ein Softwareprototyp entwickelt, der weitestgehend selbständig die Recherche nach Lebensmittelshops und die zugehörige Datenverarbeitung übernehmen kann. Das For-

1 Lebensmittel- und Futtermittelgesetzbuch.

schungsprojekt hatte eine Laufzeit von über fünf Jahren (1. Februar 2013 bis 31. August 2018) und wurde in zwei Phasen realisiert. Dies erfolgte in Zusammenarbeit dreier Departments der Hochschule für Angewandte Wissenschaften Hamburg sowie durch fachlichen Austausch mit der Zentralstelle G@zielt. Das hatte den Vorteil, dass Fachkompetenzen aus dem Bereich Information Retrieval (Department Information), Lebensmittelrecht (Department Ökotrophologie) sowie Bild- und Textanalyse (Department Medizintechnik) gebündelt wurden und die Arbeitsprozesse im Bundesamt für Verbraucherschutz und Lebensmittelsicherheit adaptiert werden konnten, die für die Umsetzung auf inhaltlicher Ebene relevant waren.

Vorgehensweise im Projekt

Durch die Analyse der Arbeitsprozesse und die Diskussion der beteiligten Departments wurden grundlegende Entscheidungen getroffen, die als Leitfaden für die Konzeption und die Programmierung der Software dienten. Dabei wurden Vorgänge und Methoden aus der Information-Retrieval-Evaluierung, insbesondere in Bezug auf Retrievalstudien und Data Mining, ausgewählt, die für die Durchführung des Projekts relevant sind. Dies greift sowohl bei der Erstellung der Datenbasis als auch im Rahmen der Bewertungen von Webdokumenten nach lebensmittelrechtlichen Kriterien.

Bei traditionellen Retrievaltests werden Dokumentenkollektionen anhand von Suchanfragen gebildet und durch Jurorinnen und Juroren bewertet. Wir haben diese Vorgehensweise dahingehend adaptiert, dass wir ebenfalls Suchanfragen zu Anwendungsfällen definieren und diese an kommerzielle Suchmaschinen versenden (für eine genaue Übersicht zur Methodik von Retrievaltests siehe Tague-Sutcliffe (1992) und für Retrievalstudien mit Web-Suchmaschinen Hawking et al. (2001)). Die Suchergebnisse werden dabei gespeichert und automatisiert durch Machine-Learning-Verfahren bewertet. Ein weiterer Grund für diese Vorgehensweise liegt in der Umsetzbarkeit. Es wäre auch denkbar, die Daten durch ein Web-Crawling zu generieren. Dabei würden aber zum einen technische und finanziellen Hürden entstehen und zum anderen eine Spezialisierung auf bestimmte Anwendungsfälle vernachlässigt. Die Probleme bestünden bei dem Betreiben der Server als auch bei der automatisierten Analyse zur Bestimmung, ob es sich um lebensmittelrelevante Dokumente handelt. Diese Vorgänge wären sehr zeitintensiv und würden Kosten verursachen, die im Rahmen eines solchen Projekts, und auch darüber hinaus, nicht finanziert werden könnten. Unser Vorgehen steht auch unter der Annahme, dass Lebensmittelhändler ein hohes Interesse daran haben, gefunden zu werden.

Für das Machine Learning zur Bewertung der Suchergebnisse wurden Verfahren der Textklassifikation nach der Support-Vector-Methode sowie eine Bild- und Logoerkennung durch die Verwendung neuronaler Netze als zielführend identifiziert. Diese genannten Analysekomponenten werden benötigt, damit alle relevanten Aspekte zur Prüfung potenzieller Verstöße auf Webseiten untersucht werden können. Die Logoerkennung ist z. B. notwendig, um Betrugsfälle zu finden, wie zu Unrecht genutzte Biosiegel. Ein weiterer Aspekt ist das risikoorientierte Monitoring, das nicht nur für die Momentaufnahme von Händler- und Produktangeboten relevant ist. Mit Hilfe von Risikokriterien wird ein Ranking der durchgeführten Fälle erstellt, das den Mitarbeiter/die Mitarbeiterin darüber informiert, welche Recherchefälle priorisiert behandelt werden sollen.

Projektziele

Zusammengefasst wurden durch die Analyse des Projektauftrags, der Arbeitsprozesse im BVL und die Anwendung wissenschaftlicher Methoden aus der Information-Retrieval-Evaluierung und dem maschinellen Lernen folgende Ziele für das Projekt definiert:

- Optimierung der Relevanz der Suchmaschinentreffer in Bezug auf die lebensmittelrechtlichen Aspekte
- Integration von Text-, Logo- und Bilderkennungsverfahren zur erweiterten Identifizierung nicht-konformer Lebensmittel unter besonderer Berücksichtigung von speziellen Kenntlichmachungen
- Entwicklung einer Monitoring-Komponente zur regelmäßigen automatisierten und risikoorientierten Abfrage definierter Lebensmittel, Inhaltsstoffe oder Händler.

Diese Ziele sollten durch eine parallele Entwicklung eines fundierten Konzepts für die Kontrolle des Online-Handels sowie einer konkreten technischen Umsetzung erreicht werden.

Konzept zur automatisierten Kontrolle des Lebensmittelmarktes im Online-Handel

In der ersten Projektphase von AAPVL wurde das Konzept zur automatisierten Kontrolle entwickelt, das als Grund-

lage für die technische Umsetzung des Softwareprototyps diente. Dieses Konzept basiert auf den Arbeitsprozessen der Zentralstelle G@zielt im BVL, die auf drei Hauptrecherchearten beruhen (Krewinkel et al. 2016). Diese Recherchearten werden im Folgenden näher erläutert.

Hauptrecherchearten in der Onlineüberwachung

Die Onlineüberwachung des Lebensmittelmarktes im Internet wird in der Praxis anhand von drei Anwendungsfällen realisiert. Diese berücksichtigen verschiedene Kontrollaspekte, ähneln sich aber in der Vorgehensweise zur Datenbeschaffung. Die Hauptrecherchearten sind dabei folgendermaßen gestaltet:

- Unternehmensrecherche: Bei der Unternehmensrecherche handelt es sich um die Überprüfung der Registrierungspflicht von Lebensmittelhändlern nach Art. 6 Abs. 2 der europäischen Lebensmittelhygiene-Verordnung (EG) Nr. 852/2004. Hiernach muss sich jeder Lebensmittelhändler bei der zuständigen Behörde anmelden. Online-Lebensmittelhändler bilden dabei keine Ausnahme.
- Anbieterrecherche: Bei einer Anbieterrecherche liegt der Fokus auf der Suche nach möglichst allen Anbietern einer bestimmten Produktkategorie. Dadurch sollen die Behörden sich einen Überblick zu dem Händlerangebot für bestimmte Produkte verschaffen oder kritische Produktgruppen gezielt im Blick behalten können.
- Produktrecherche: Bei der Produktrecherche handelt es sich um eine risikoorientierte Recherche nach bereits bekannten nicht verkehrsfähigen Produkten. Sie dient der Ermittlung von Anbietern, die Produkte in ihrem Onlineangebot haben, die bereits als nicht verkehrsfähig eingestuft wurden.

Durch die Definition der Abläufe und Ziele dieser Recherchearten sowie die Ermittlung der Gemeinsamkeiten konnten die einzelnen Prozesse abstrahiert werden, die für die Entwicklung des Softwareprototyps ausschlaggebend waren. So besteht die Aufgabe des Tools hauptsächlich darin, Webseiten zu analysieren und basierend auf den Analyseergebnissen zu filtern, so dass am Ende eine relevante und handhabbare Datenmenge generiert wird.

Technisches Konzept für den Software-Prototyp

Anhand der Analyse der Recherchearten konnten die einzelnen Abläufe und Module für das Konzept entwickelt werden. Das Ergebnis ist ein mehrstufiger Prozess mit verschiedenen Filterstufen, um den Mitarbeiterinnen und Mitarbeitern der Behörden handhabbare Datenmengen an potenziellen Verstößen gegen das Lebensmittelrecht zur weiteren rechtlichen Prüfung und ggf. zur Durchführung von behördlichen Maßnahmen zur Verfügung zu stellen (s. Abb. 1.)

In der ersten Filterstufe wird die Datenerfassung mit Hilfe von Suchanfragen umgesetzt. Diese werden an verschiedene Suchmaschinen geschickt und die Suchergebnisse in Form von Screenshots sowie dem zugehörigen Quelltext gespeichert. Dabei werden Duplikate herausgefiltert sowie die Domains mit einer Negativliste abgeglichen, um sicherzustellen, dass Websites, die keinen Bezug zur Online-Vermarktung von Lebensmitteln haben wie z. B. Wikipedia, von den Analysen ausgeschlossen werden.

Der nächste Schritt der Analyse ist die Einteilung der Ergebnisse in Online-Shops und in spezielle Lebensmittel-Shops, da im Kontext der Lebensmittelkontrolle nur Websites relevant sind, die auch tatsächlich Lebensmittel anbieten. Darauffolgend findet eine weitere Klassifizierung statt, die sicherstellen soll, dass es sich bei dem Ergebnis um ein Produktangebot handelt. Dies ist vor allem für die Verdachtsanalyse wichtig, da dann konkret Informationen wie Zutatenverzeichnisse oder Artikelbezeichnungen ausgewertet werden können. Bei der Verdachtsanalyse werden durch Text- und Bilderkennung die Informationen gewonnen, die für einen Abgleich mit hinterlegten lebensmittelrechtlichen Kriterien genutzt werden. Auf Basis dieser Auswertung werden potenzielle Verstöße ermittelt und für eine manuelle Nachbewertung durch Mitarbeiterinnen und Mitarbeiter der Überwachungsbehörden aufbereitet. Um die automatisiert ermittelten Ergebnisse der Vor-Ort-Kontrolle zuführen zu können, findet eine automatisierte Impressumserkennung mit anschließender Zuordnung zu Bundesland und Kreis statt.

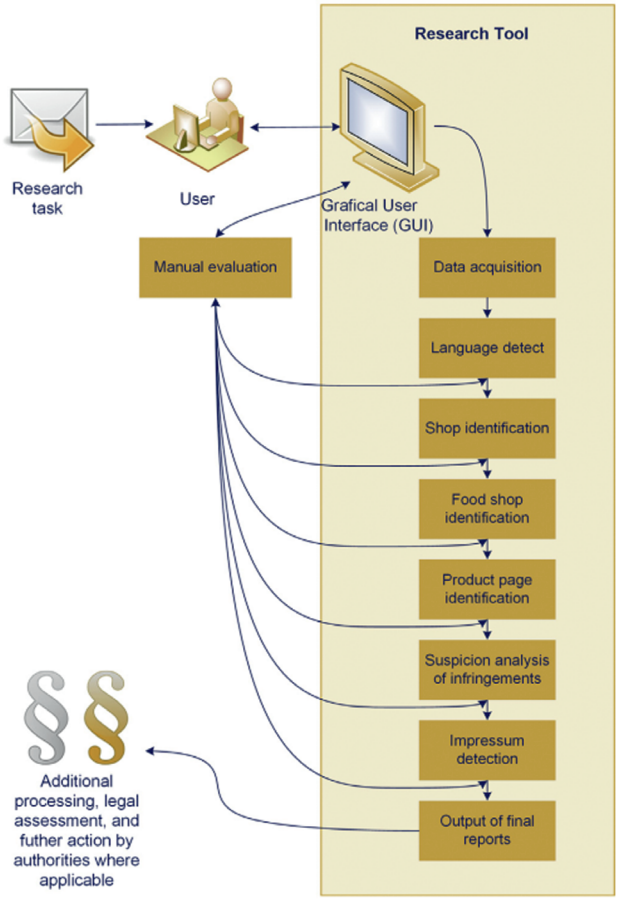

Abbildung 1: Konzept zur Überwachung des Onlinehandels (Krewinkel et al. 2016, S. 206).

Erweiterung des Konzepts durch Elemente für das risikoorientierte Monitoring

Die zweite Projektphase hatte das Ziel, zu dem ursprünglichen Konzept zur Kontrolle des Lebensmittelmarkts weitere Abläufe und Kriterien zu definieren, die eine risikoorientierte Überwachung ermöglichen. Dazu wurden inhaltliche Kriterien analog zu den Anforderungen an die risikobasierte Überwachung im stationären Handel erarbeitet und die relevanten Unterschiede herausgearbeitet, um diese für die Überwachung im Onlinehandel zu adaptieren.

Das Risiko wird jeweils berechnet, indem eine Kombination von vorab definierten Risikokriterien mit den Analyseergebnissen der Software sowie manuellen Nachbewertungen durch die Mitarbeiterinnen und Mitarbeiter der Überwachungsbehörden kombiniert werden. Dies führt schließlich zu einem Ranking der durchgeführten Recherchen in einer absteigenden Liste nach den ermittelten Risikopunkten. So lässt sich ableiten, welche Recherchen eine hohe Priorität aufweisen und durch die jeweiligen Bearbeiter wiederholt werden müssen.

In dem Forschungsprozess zur Definition der Kriterien zur Ermittlung der Risikopunkte wurden die relevanten Quellen und Gesetzestexte wie die AVV Rüb[2], in der z. B. eine Mindest- und Höchstkontrollhäufigkeit vorgeschrieben wird, berücksichtigt. Laut § 6(2) AVV Rüb beträgt diese Kontrollfrequenz höchstens täglich bis in der Regel mindestens alle drei Jahre. Dies ist einer der wichtigsten Ausgangspukte für die Sortierung der Fälle in der Software. Denn entsprechend dieser Vorschrift ist im Tool sichergestellt, wie häufig Kontrollen durchgeführt werden sollen. Dabei gibt es allerdings den Unterschied, dass nicht die Anbieter selbst kontrolliert werden, sondern auf Basis der durchgeführten Recherchen Kontrollfrequenzen ermittelt werden. Je nach vergangener Zeit werden die Punkte folgendermaßen vergeben, um die Priorisierung einer Nachrecherche zu verändern:

- Höchste Priorität – Recherchen, die >36 Monate her sind (2 Punkte)
- Mittlere Priorität – Recherchen innerhalb der letzten 13 bis 36 Monate (1 Punkt)
- Geringste Priorität – Recherchen innerhalb der letzten 12 Monate (0 Punkte)

Neben der Entwicklung der Kriterien anhand der Verwaltungsvorschriften wurden die Filterkriterien aus dem europäischen Frühwarnsystem RASFF[3] für die Vorabvergabe von Risikopunkten für die Recherchefälle adaptiert. Innerhalb des Portals im RASFF lassen sich Warnmeldungen anhand von verschiedenen Kriterien filtern. Dazu zählen die Art des Falls (Futtermittel, Lebensmittel oder Lebensmittel-Kontaktmaterial), die Risikoeinschätzung (bedenklich, unbedenklich, unbestimmt) sowie die Klassifikation des Falls in die Gruppen Alarm, Grenzfall, Information, Kenntnisnahme und Weiterverfolgung. Diese Unterteilung hat bereits einen wichtigen Anteil an der Risikoklassifizierung. Zusätzlich zu der Berechnung des Scores können noch manuell Punkte vergeben werden,

2 Allgemeine Verwaltungsvorschrift über Grundsätze zur Durchführung der amtlichen Überwachung der Einhaltung der Vorschriften des Lebensmittelrechts, des Rechts der tierischen Nebenprodukte, des Weinrechts, des Futtermittelrechts und des Tabakrechts: http://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_03062008_3158100140002.htm [1.12.2018].

3 RASFF – the Rapid Alert System for Food and Feed: https://ec.europa.eu/food/safety/rasff/portal_en [1.12.2018].

da die vorgestellten Kriterien nicht allen Rechercheaufträgen gerecht werden.

Aufbau und Technik des Softwareprototyps

Der Softwareprototyp stellt die konkrete technische Umsetzung des vorgestellten Konzepts mit der Erweiterung der risikoorientierten Analysen dar und wird im Folgenden näher erläutert. Der Software-Prototyp ist eine Web-Anwendung, die auf einem Webserver installiert wird und durch einen Webbrowser nutzbar ist. Dabei können die Anwender auswählen, welche Art der Recherche gestartet werden soll (s. o.). So findet keine Verdachtsanalyse statt, wenn eine Unternehmensrecherche gestartet wird, da diese nur das Ziel verfolgt, ein möglichst vollständiges Bild an Lebensmittelhändlern im Web zu liefern. Für diesen Zweck wurden auch ca. 200 generische Suchanfragen im Tool hinterlegt.

Generierung von Suchanfragen in der Software

Der Software-Prototyp bietet die Möglichkeit, Vorlagen für Suchanfragen zu hinterlegen, die durch den Anwender zur Anreicherung oder Generierung von Suchanfragen nutzbar ist. So müssen beispielsweise bei den Unternehmensrecherchen, die immer auf den gleichen Suchanfragen basieren, nicht jedes Mal diese Anfragen selbst eingegeben werden. Hierfür sind Vorlagen entstanden, die aus Suchanfragen aus einer Lebensmittelkategorie und Verben bestehen, die eine Kaufabsicht anzeigen. Die Lebensmittelkategorien wurden dabei der Lebensmittelpyramide der Deutschen Gesellschaft für Ernährung entnommen (DGE, 2016). Beispiele für solche Suchanfragen sind „Bier online bestellen“, „Lebensmittel liefern lassen“ oder „Kaffee kaufen“.

Softwarearchitektur des Prototyps

Für die Umsetzung der Software wurde auf das verbreitete Model-View-Controller-Prinzip (MVC) zurückgegriffen (s. Abb. 2). Bei diesem Programmiermuster werden die einzelnen Komponenten und Module einer Software verschiedenen Schichten zugeordnet. Damit wird eine gute Wartbarkeit erzielt und Möglichkeiten für die Erweiterung geschaffen, da sich die Komponenten austauschen und in die Software integrieren lassen[4] (Eckerson, 1995 & Fowler, 2002). So wäre es z. B. denkbar, die Anwendungen in der Präsentationschicht vollkommen auszutauschen, um die grafische Benutzerschnittstelle in einer anderen Programmiersprache zur Verfügung zu stellen oder auch die Analysewerkzeuge mit anderen Programmbibliotheken und Schnittstellen zu erweitern, ohne die übrigen Module anpassen zu müssen.

Die technische Umsetzung erfolgte anhand geeigneter Programmiersprachen für den Anwendungsfall. Da die Software webbasiert ist und in jedem herkömmlichen Browser nutzbar sein sollte, wurde die Benutzeroberfläche in PHP und Javascript entwickelt. Beides sind De-facto-Standards für die Entwicklung von Webanwendungen. Für die Analysekomponenten wurde die Programmiersprache Python gewählt, da diese beispielsweise mit der Programmbibliothek SciPy[5] und scikit-learn[6] hervorragende Tools für Datenanalysen und Machine Learning bereitstellt (Garreta, R., & Moncecchi, G., 2013).

Für das AAPVL-Tool verwendete man auch in anderen Forschungsprojekten entwickelte Programmmodule. So wurde die Technologie für das Abfragen der Suchmaschinen mit einem Suchmaschinenscraper aus dem *Relevance Assessment Tool* (RAT) (Lewandowski & Sünkler, s. auch den Artikel zum RAT in diesem Heft) übernommen und für den Softwareprototyp erweitert. *Screen scraping* ist allgemein eine Methode, um Daten aus Dokumenten anhand von Merkmalen zu extrahieren. In der Anwendung im Relevance Assessment Tool und in der AAPVL-Software bedeutet dies, die Suchergebnisseiten (*search engine result pages*, SERPs) von Suchmaschinen anhand von Suchanfragen automatisiert auszulesen. Dabei wird der HTML-Quelltext der Seiten eingelesen. Anhand von Tags zur Strukturierung und Formatierung des Quelltexts werden die relevanten Informationen, wie die URL der Suchergebnisse und die Trefferbeschreibung, gespeichert. Mit diesem Ansatz kann jeweils die maximale Anzahl an zurückgegebenen Suchergebnissen zu einer Anfrage mit den Quelltexten und Screenshots der Webseiten für weitere Analyseprozesse gespeichert und auf der Text- und der Bildebene verarbeitet werden.

4 Die Software ist Open Source und kann in dem Repository https://bitbucket.org/ssuenkler/aapvl/src/master/[1.12.2018] bezogen werden.
5 https://www.scipy.org/ [1.12.2018].
6 https://scikit-learn.org/ [1.12.2018].

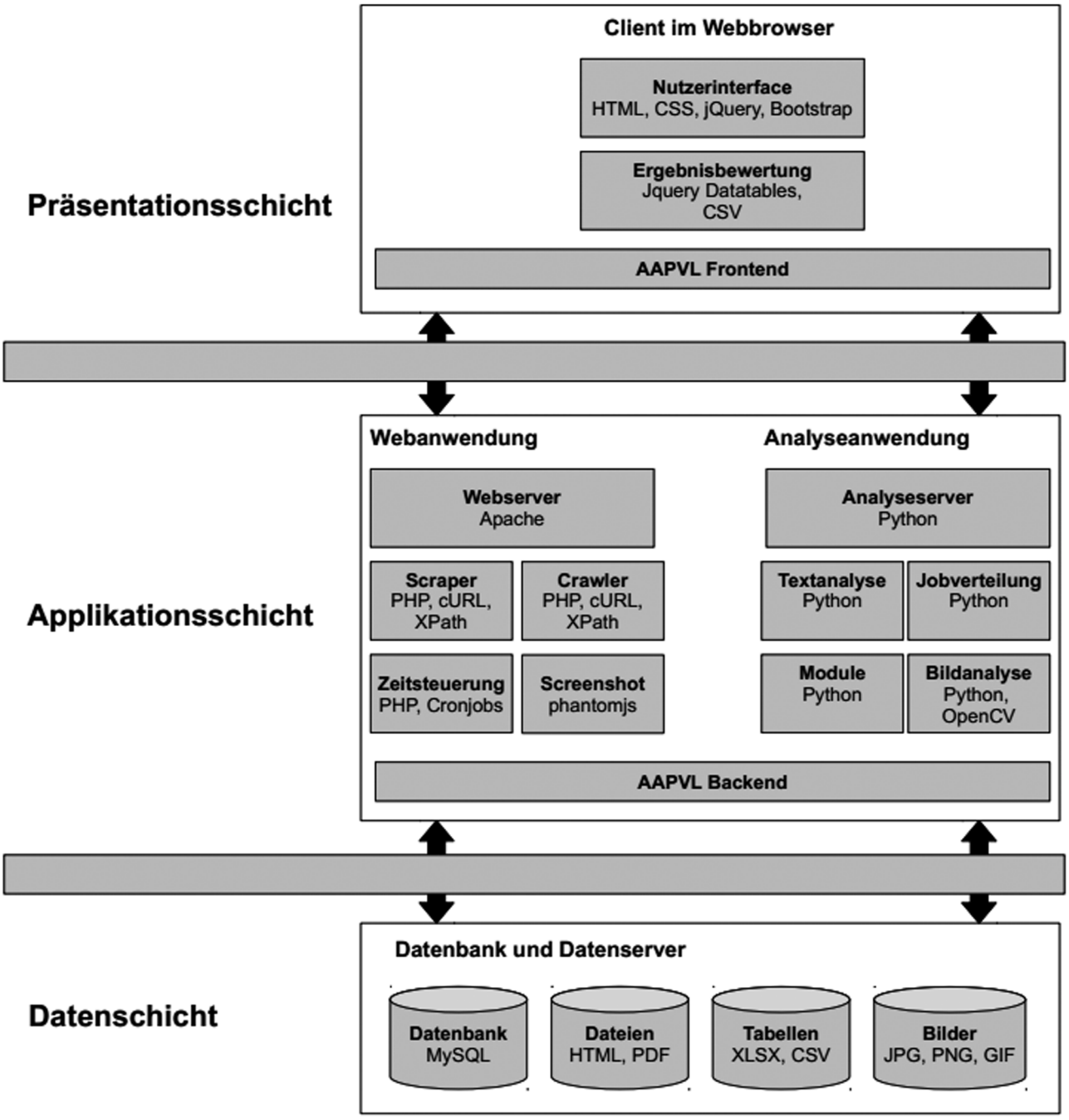

Abbildung 2: Softwarearchitektur des Prototyps zur Entwicklung von automatisierten Analyseverfahren zur Identifizierung und Bewertung von nicht verkehrsfähigen Produkten des virtuellen Lebensmittelmarktes.

Umsetzung der MVC-Architektur

Wie bereits erwähnt, wurde für die Entwicklung des Tools eine Model-View-Controller-Architektur gewählt. Dabei stehen das Model für die Datenschicht, die View für die Präsentationsschicht und der Controller für die Applikationsschicht. Die Schichten werden im Folgenden mit ihren jeweiligen Modulen und Aufgaben näher erläutert.

Datenschicht

Sämtliche Module in der Anwendungsebene kommunizieren über eine MySQL-Datenbank. Dort werden sowohl die Ergebnisse aus der Datenerfassung (gespeicherte Webseiten und Bilder) als auch sämtliche Analyseergebnisse sowie die Eingaben und Metadaten zu den Recherchefällen gespeichert und nach Aufbereitung durch SQL-Abfragen

auf der Präsentationsebene dem Nutzer verfügbar gemacht. Da die Anwendung in zwei Programmiersprachen entwickelt wurde, dient die MySQL-Datenbank auch als wichtige Schnittstelle zwischen den Modulen in den verschiedenen Sprachen.

Applikationsschicht

In der Applikationsschicht werden sämtliche Eingaben verarbeitet und die Suchergebnisse mit einem Scraper erfasst sowie die Impressen durch einen Crawler auf den Webseiten ausgelesen. Die gespeicherten URLs werden im nächsten Schritt aufgerufen, um den Quelltext zu sichern und einen Screenshot als Bild anzufertigen. Der Crawler hingegen durchsucht die aufgerufene URL nach der Unterseite mit den Kontaktdaten auf der Webseite bzw. nach dem Impressum. Dies geschieht durch den Abgleich des Hyperlinks und des Linktextes mit Keywords wie Impressum, Kontakt, Über uns, usw.

Scraper und Crawler wurden dabei in PHP mit der Nutzung von PhantomJS programmiert. PhantomJS ist ein sogenannter Headless Browser und simuliert Anfragen an Webseiten in der Rolle herkömmlicher Webbrowser. Wenn diese Daten erfasst sind, werden damit automatisch sogenannte Jobs generiert, die durch die Analyseanwendung aufgegriffen und verarbeitet werden. Es finden pro gespeicherte Webseite Klassifikationen statt, die auf Algorithmen aus dem überwachten maschinellen Lernen basieren. Durch den Einsatz einer Support Vector Machine werden die Inhalte im Quelltext gegen eine zuvor erstellte Wissensbasis abgeglichen und eine Wahrscheinlichkeit für die Übereinstimmung berechnet. Diese Vorgänge sind in Python implementiert, da diese Sprache einen stärkeren Fokus auf die Datenanalyse hat.

Für das AAPVL-Tool wurden über 1.000 Webseiten für die Shop-Erkennung und die Lebensmittelshop-Erkennung manuell klassifiziert; sie bilden die Wissensbasis. Daraus wurden automatisch die Features extrahiert, die als Indikatoren für die Einschätzung neuer Webseiten in Bezug auf die Eigenschaften von Online-Shops und Lebensmittel-Shops herangezogen werden.

Je nach Recherche-Art folgen noch weitere Schritte. So wurde eine Verdachtsanalyse für die Überprüfung des Logo- und Missbrauchs von Biosiegeln implementiert. Dafür wird auf den Webseiten bzw. auf den Screenshots über eine Bilderkennung ausgewertet, ob sich dort ein Biosiegel befindet. Wenn dies der Fall ist, wird die so genannte Öko-Kontrollnummer aus dem Quelltext extrahiert und mit einer Liste zulässiger Nummern abgeglichen. Fällt diese Überprüfung negativ aus, wird das Web-Ergebnis entsprechend gekennzeichnet.

Präsentationsschicht

Die ausgewählten Technologien in der Präsentationsschicht umfassen eine Kombination von JavaScript-Bibliotheken mit HTML und CSS. Diese Kombination gilt als Defacto-Standard für die modernen Web-Applikationen. Die Hauptkomponente der Präsentationsschicht ist die grafische Benutzeroberfläche (GUI) oder auch das Frontend, das verantwortlich für die Anzeige der Ergebnisse aus der Applikationsschicht ist und auch die Eingaben in verschiedenen Formularen verarbeitet. Mit Hilfe von Konfigurationsdateien können auch Sprachtemplates definiert werden, damit eine Nutzung in weiteren Staaten ermöglicht wird.

Neben der Möglichkeit, die Daten für die Recherchearten sowie die Suchanfragen dafür festzulegen, lassen sich im Frontend die Ergebnisse der Recherchen manuell nachbewerten, da die automatisierten Analysen keine hundertprozentig verlässlichen Ergebnisse liefern können. Damit lassen sich falsch klassifizierte Inhalte korrigieren sowie Verdachtsmomente, die durch die Software identifiziert wurden, verifizieren. Dies erfolgt in Form eines Editors, der eine Bearbeitung der Inhalte in Tabellen ermöglicht. Abbildung 3 zeigt einen Ausschnitt dieses Editors.

Erkennungsraten der Klassifikatoren Shop- und Lebensmittelshoperkennung

Die Erkennungsraten der Klassifikatoren wurden in einer umfangreichen Unternehmensrecherche getestet. Es wurden insgesamt 121 Suchanfragen an google.de gesendet und die Ergebnisse automatisiert gespeichert. Jedes Suchmaschinenergebnis wurde mittels Crawling der entsprechenden Hauptdomain zugeordnet sowie die zugehörige Impressumsseite ermittelt. Insgesamt kamen 137.619 Suchergebnisse zustande, die sich auf 39.452 Domains verteilen. Diese Seiten wurden in die Klassen „Shop“ und „Lebensmittelshop“ klassifiziert. Eine manuelle Auswertung von 11.115 Seiten in Gegenüberstellung zu den Ergebnissen der maschinellen Klassifikatoren auf diesen Seiten zeigt die Präzision der Klassifikation.

Mit dieser Studie sollte der Schwellenwert für die Wahrscheinlichkeit bestimmt werden, ab wann eine Webseite als Shop oder kein Shop bzw. Lebensmittelshop oder

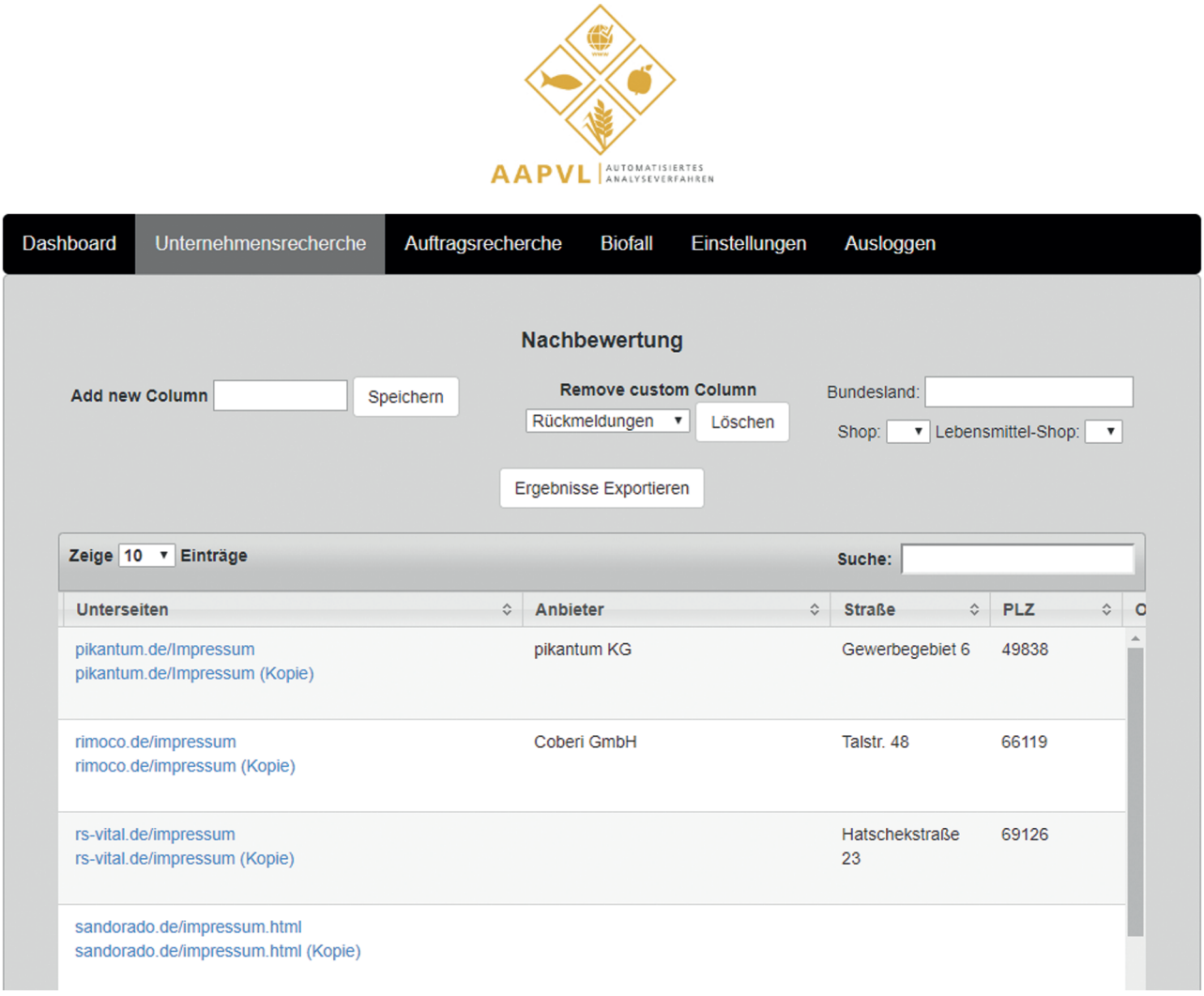

Abbildung 3: Editor zur Nachbewertung in der AAPVL-Anwendung.

kein Lebensmittelshop klassifiziert wird. Das Ziel war, das beste Verhältnis zwischen falschen und richtigen Zuordnungen zu ermitteln. Hierbei galt, dass die Richtig-positiv-Rate möglichst hoch, wichtiger aber die Falsch-negativ-Rate so niedrig wie möglich sein soll, damit keine relevanten Lebensmittelhändler der Kontrolle entgehen. Um den Behörden unnötigen Aufwand zu ersparen, ist ein möglichst geringer Wert der Falsch-positiv-Rate nötig. Für das Herausfiltern entsprechend vieler irrelevanter Ergebnisse muss der Wert der Richtig-negativ-Rate möglichst hoch sein. Abbildung 4 illustriert die Verteilung der Ergebnisse des Shop-Klassifikators auf die vier Ergebnisraten bei den entsprechenden Sicherheitswerten. Abbildung 5 zeigt die entsprechende Grafik für den Lebensmittelshop-Klassifikator.

Es ist ersichtlich, dass der Klassifikator für die Shoperkennung deutlich zuverlässiger klassifiziert als der Lebensmittelshop-Klassifikator. Daher wurde für die Shoperkennung der optimale Wert bei der Erkennungsgrenze von 85 Prozent und bei der Lebensmittelshoperkennung bei einem Wert von 80 Prozent gesetzt. Der Prototyp ist also so voreingestellt, dass alle Seiten, die mit einer Wahrscheinlichkeit von unter 85 Prozent kein Shop sind, und nur die Seiten, die mit einer Wahrscheinlichkeit von 85 Prozent oder mehr eingestuft wurden, in die weitere Bearbeitung (z. B. manuelle Nachbewertung) gehen. Gleiches gilt bei der Lebensmittelshoperkennung bei einem Wert von 80 Prozent. Diese Voreinstellung lässt sich jedoch in Abhängigkeit von der Fragestellung oder der vorhandenen Bearbeitungszeit entsprechend der individuellen Bedürfnisse fallspezifisch anpassen.

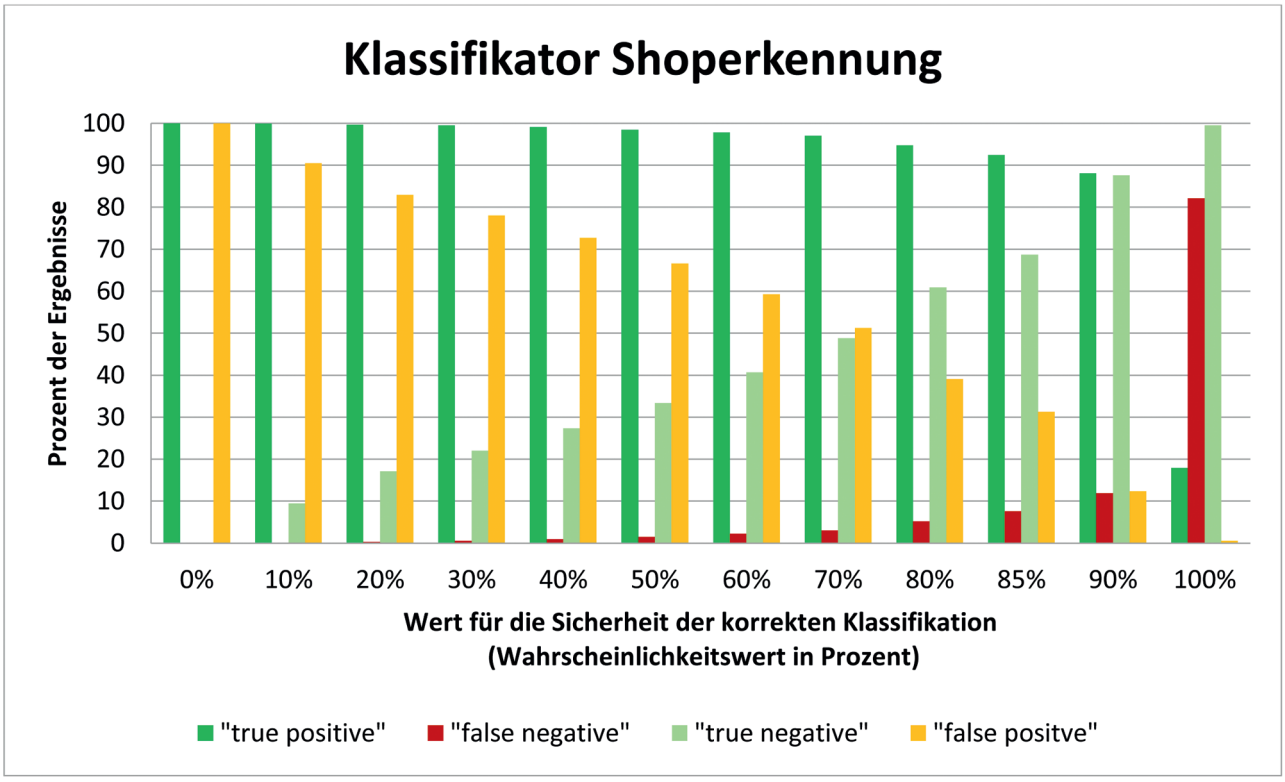

Abbildung 4: Zuverlässigkeit des Klassifikators Shoperkennung in Abhängigkeit zum Wahrscheinlichkeitswert für eine korrekte Klassifikation.

Fazit

Das hier vorgestellte Forschungsprojekt bietet einen wichtigen Beitrag zur Lebensmittelüberwachung bei der Kontrolle des Online-Handels mit Lebensmitteln, da die Schnelllebigkeit und die unüberschaubare Menge an Daten große Herausforderungen darstellen. Mit dem entwickelten Konzept und dem dazugehörigen Softwareprototyp lassen sich bereits jetzt wiederkehrende Tätigkeiten zur Durchführung von Unternehmensrecherchen, Produktrecherchen und Auftragsrecherchen unterstützen. Die Software wurde im Rahmen verschiedener Auftragsrecherchen durch das Bundesamt für Verbraucherschutz und Lebensmittelsicherheit genutzt und hat sich dort für die Unternehmensrecherche und die Produktrecherche bewährt. Ein Fall zur Überprüfung der rechtmäßigen Nutzung von Biosiegel ist noch in der Bearbeitung und wird nach Fertigstellung ausgewertet.

Im Rahmen des Projektes wurden nicht nur wichtige Erkenntnisse und Ergebnisse für die Lebensmittelüberwachung generiert. Auch für die Informationswissenschaft bilden solche langjährigen Verbundprojekte große Chancen, da dort gezeigt werden kann, dass auch solche speziellen Anwendungsfälle informationswissenschaftliche Kompetenzen benötigen. So wurden insbesondere Methoden aus dem Information Retrieval und der Information-Retrieval-Evaluierung zur Gestaltung des Konzepts und für die technische Umsetzung angewendet. Bei Retrievaltests werden Dokumentenkollektionen anhand von Suchanfragen generiert und durch Jurorinnen und Juroren bewertet. Dabei werden auch die Relevanzkriterien als Bewertungsgrundlage festgelegt. Im AAPVL-Projekt ist die Vorgehensweise analog, wobei die Bewertung hauptsächlich durch Machine-Learning-Konzepte in der Software stattfindet.

Die vorgestellte Software und ihre Komponenten sind sehr flexibel gestaltet und nicht auf ihre Anwendung im Lebensmittelkontext beschränkt. Das Scrapen von Suchmaschinen für die Zusammenstellung von Dokumentenkollektionen lässt sich für andere Anwendungen einsetzen. Die Klassifikationskomponenten sind dazu in der Lage, jede Art von Webdokument nach Vorgaben zu klassifizieren. So können die Klassifkationskomponenten für jede Art von Webseitenklassifizierung nach bestimmten Themen und Kriterien genutzt werden, z. B. im Rahmen von Marktsichtungen oder zur Erkennung von Produktfälschungen. Dafür wird nur eine Wissensbasis benötigt, die vorklassifizierte Webseiten in Form von Quelltexten enthält. Analog lässt sich die Bilderkennung auch für die Analyse anderer Logos oder Bilder einsetzen. Auch dafür

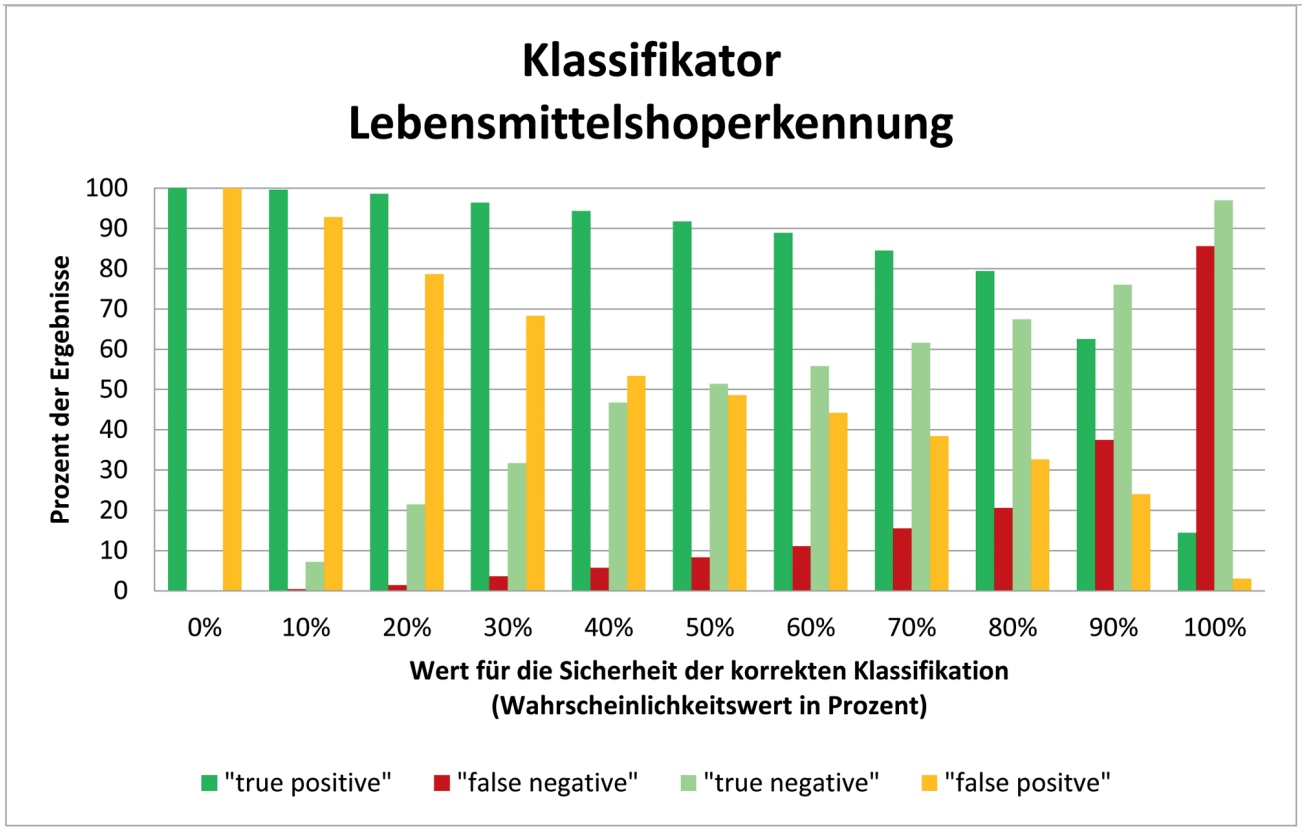

Abbildung 5: Zuverlässigkeit des Klassifikators Lebensmittelshoperkennung in Abhängigkeit zum Wahrscheinlichkeitswert für eine korrekte Klassifikation.

müssen Trainingsdaten aufbereitet und bereitgestellt werden. Die Impressumserkennung ist ebenfalls eine generische Anwendung, die z. B. für die Zusammenstellung von Kontaktdaten nutzbar ist, was wiederum für eine Marktsichtung hilfreich ist.

Zusammenfassend betrachtet, ist dieses Projekt nur ein Beispiel dafür, wie sich mit grundlegenden Erkenntnissen aus der Informationswissenschaft Klassifikations- und Analyseaufgaben zu einer Vielzahl von Anwendungsfällen durchführen lassen, wenn Webdokumente als Dokumentenbasis automatisiert ausgewertet werden sollen. Es ergeben sich dadurch zahlreiche Chancen, auch in Bereichen abseits von traditionellen Information-Retrieval-Evaluierungen oder automatischen Textklassifikationen. Ferner zeigt die Durchführung eines solchen Projekts, dass die Softwareentwicklung bei der Nutzung vorhandener Programmbibliotheken und bewährter Konzepte mit überschaubarem Aufwand betrieben werden kann. So existieren sowohl für die Gestaltung von Weboberflächen zahlreiche Open Source-Frameworks[7] und für das Machine-Learning bietet Python auch sehr gute Open Source-Bibliotheken[8].

Förderhinweis

Die Förderung des Projekts erfolgte aus Mitteln des Bundesministeriums für Ernährung und Landwirtschaft (BMEL) aufgrund eines Beschlusses des deutschen Bundestages. Die Projektträgerschaft erfolgt über die Bundesanstalt für Landwirtschaft und Ernährung (BLE) im Rahmen des Programms zur Innovationsförderung. Das Projekt hatte die Förderkennzahl 2819104715.

7 z. B. ist AngularJS ein Open Source Framework von Google, das alle notwendigen Funktionen zur Entwicklung von Webapplikationen mitbringt.

8 Dazu gehört u. a. die scikit-learn-Bibliothek, die für eine Vielzahl von Machine-Learning-Algorithmen entwickelt wurde.

Literatur

Bundesverband E-Commerce und Versandhandel Deutschland e.V. (bevh). (2018) . *E-Commerce – der neue Nahversorger?* https://cloud.bevh.org/index.php/s/bVmooV05I64DkQD/download?path=%2F&files=180122 %20bevh_Praesentation%20E-Commerce%20der%20neue%20Nahversorger%3F.pdf [1.12.2018].

BVL. (o. J.). G@ZIELT. Sich im Internet einkaufen. https://www.bvl.bund.de/DE/01_Lebensmittel/01_Aufgaben/06_UeberwachungInternethandel/lm_ueberwachung_internethandel_node.html [1.12.2018].

Deutsche Gesellschaft für Ernährung e. V. (DGE). (2016). *Dreidimensionale DGE-Lebensmittelpyramide*. https://www.dge-medienservice.de/fileuploader/download/download/?d=0&file=custom%2Fupload%2FFile-1469015023.pdf [1.12.2018].

Eckerson, W. (1995). Three tier Client/Server architecture: achieving scalability, performance, and efficiency in client server applications. Open Information Systems, 10(1).

Fowler, M. (2002). Patterns of enterprise application architecture. Boston: Addison Wesley.

Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn. Machine Learning in Python*. Packt Publishing Ltd., UK.Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring Search Engine Quality. Information Retrieval, 4(1), 33–59.

IfD. (2018). *Allensbacher Markt- und Werbeträger-Analys*. AWA 2018.

Krewinkel, A., Sünkler, S., Lewandowski, D., Finck, N., Tolg, B. Kroh, L. W., Schreiber, G. A., Fritsche, J. (2016). Concept for automated computer-aided identification and evaluation of potentially non-compliant food products traded via electronic commerce. *Food Control 61 (2016,)* 204–212. http://dx.doi.org/10.1016/j.foodcont.2015.09.039.

Lewandowski, D., & Sünkler, S. (2013). Designing search engine retrieval effectiveness tests with RAT. *Information Services & Use*, *33*(1), 53–59.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. Information Processing & Management, 28(4), 467–490.

Prof. Dr. Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät DMI, Department Information
Finkenau 35
22081 Hamburg
dirk.lewandowski@haw-hamburg.de

Dirk Lewandowski ist Professor für Information Research & Information Retrieval an der Hochschule für Angewandte Wissenschaften Hamburg. Seine Forschungsinteressen sind Web Information Retrieval, Qualitätsfaktoren von Suchmaschinen, das Rechercheverhalten der Suchmaschinen-Nutzer sowie die gesellschaftlichen Auswirkungen des Umgangs mit den Web-Suchmaschinen.www.searchstudies.org/dirk

Alexandra Krewinkel
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Life Sciences, Department Ökotrophologie
Ulmenliet 20
21033 Hamburg
alexandra.krewinkel@haw-hamburg.de

Alexandra Krewinkel hat einen Bachelor in Ökotrophologie an der Justus Liebig Universität in Gießen absolviert und wechselte daraufhin für den Master in Food Science an die Hochschule für Angewandte Wissenschaften in Hamburg. Sie arbeitet als Referentin im Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL) bei der Zentralstelle G@zielt, die den Online-Lebensmittelmarkt in Deutschland kontrolliert und war Teil des Forschungsprojekts AAPVL.

Mareike Gleissner
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Life Sciences, Department Ökotrophologie
Ulmenliet 20
21033 Hamburg
mareike.gleissner@haw-hamburg.de

Mareike Gleissner studierte Ökotrophologie (M.Sc.) an der Christian-Albrechts-Universität zu Kiel und Drug Regulatory Affairs (M.D.R.A.) an der Rheinischen Friedrich-Wilhelms-Universität Bonn. Als wissenschaftliche Mitarbeiterin der Hochschule für Angewandte Wissenschaften Hamburg arbeitete sie in einem Projekt zur Entwicklung automatisierter, risikoorientierter Kontrollen im Online-Lebensmittelhandel. Im Rahmen ihrer weiteren beruflichen Tätigkeiten in der Lebensmittelindustrie beschäftigte sie sich insbesondere mit der Entwicklung und Vermarktung spezieller, funktioneller Lebensmittel.

Dorle Osterode
Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Life Sciences, Department Ökotrophologie
Ulmenliet 20
21033 Hamburg
dorle.osterode@haw-hamburg.de

Dorle Osterode hat ihren Master of Science in Informatik und hat in dem Projekt AAPVL die Software für die Analyseprozesse zur Text- und Bilderkennung entwickelt.

Prof. Dr.-Ing. Boris Tolg
Hochschule für Angewandte
Wissenschaften Hamburg
Fakultät Life Sciences, Department
Ökotrophologie
Ulmenliet 20
21033 Hamburg
boris.tolg@haw-hamburg.de

Boris Tolg ist seit 2008 Professor für Informatik und Mathematik an der Hochschule für Angewandte Wissenschaften Hamburg. In den Jahren davor war er Softwarearchitekt bei einem großen deutschen Automobilzulieferer und entwickelte Infotainmentgeräte. Seine Forschungsinteressen liegen im Bereich der Simulation (in der realen und der virtuellen Realität) und Analyse von Großschadenslagen z. B. durch Bewegungsanalysen. Er ist Leiter des SIMLab an der HAW Hamburg

Prof. Dr. Martin Holle
Hochschule für Angewandte
Wissenschaften Hamburg
Fakultät Life Sciences, Department
Ökotrophologie
Ulmenliet 20
21033 Hamburg
martin.holle@haw-hamburg.de

Martin Holle ist Professor für Lebensmittelrecht und Allgemeines Verwaltungsrecht an der Hochschule für Angewandte Wissenschaften in Hamburg. Zuvor war er zwölf Jahre lang als Rechtsanwalt in einem internationalen Lebensmittelunternehmen in Hamburg, Rotterdam und London tätig. Er ist Mitglied der Deutschen Lebensmittelbuch-Kommission und des Wissenschaftlichen Beirats der Wissenschaftlichen Gesellschaft für Lebensmittelrecht.

Sebastian Sünkler
Hochschule für Angewandte
Wissenschaften Hamburg
Fakultät DMI, Department Information
Finkenau 35
22081 Hamburg
sebastian.suenkler@haw-hamburg.de

Sebastian Sünkler ist wissenschaftlicher Mitarbeiter und Promovierender an der Hochschule für Angewandte Wissenschaften Hamburg. Er ist Teil der Forschungsgruppe Search Studies und hat dort von Beginn an der Entwicklung des Relevance Assessment mitgewirkt. In den letzten Jahren hat er in dem Projekt AAPVL gearbeitet, in dem eine Software zur automatisierten Kontrolle des Lebensmittelmarktes entwickelt wurde. Seine Forschungsinteressen sind die Evaluierung von Suchmaschinen und dialogbasierten intelligenten Assistenten. www.searchstudies.org/sebastian