Food Control 61 (2016) 204–212

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Concept for automated computer-aided identification and evaluation of potentially non-compliant food products traded via electronic commerce

Alexandra Krewinkel [a], Sebastian Sünkler [b], Dirk Lewandowski [b], Niklas Finck [a], Boris Tolg [a], Lothar W. Kroh [c], Georg A. Schreiber [d, 1], Jan Fritsche [a, *]

[a] *Hamburg University of Applied Sciences, Faculty Life Sciences, Ulmenliet 20, 21033, Hamburg, Germany*
[b] *Hamburg University of Applied Sciences, Faculty Design, Media, Information, Finkenau 35, 22081, Hamburg, Germany*
[c] *Institute of Food Technology and Food Chemistry, Berlin University of Technology, Gustav-Meyer-Allee 25, 13355, Berlin, Germany*
[d] *Federal Office of Consumer Protection and Food Safety, Mauerstraße 39 – 42, 10117, Berlin, Germany*

ARTICLE INFO

Article history:
Received 12 August 2015
Accepted 27 September 2015
Available online 30 September 2015

Keywords:
e-commerce
Internet
e-food
Food control
Food law enforcement
Automatization

ABSTRACT

The online marketplace for food products is continually expanding and all types of food and beverages can now be purchased over the internet. It is primarily the responsibility of the food business operator to ensure compliance with food safety law. However, the competent authorities are tasked with controlling the e-food sector as part of their regulatory duties to protect consumer health and to prevent fraud, regardless of the sales channel being used. For this purpose, a new software prototype concept was developed that automatically identifies and evaluates potentially non-compliant e-food products. The prototype was developed using a modular architecture comprising a research tool, an image analysis tool, and a monitoring tool. User-defined thresholds are applied to assess the reliability of the retrieved data. Results that are not deemed reliable enough can be reworked using a computer-aided evaluation interface. The research tool utilizes both internet search engines and customized search algorithms. A multi-stage filtering process is performed to limit the sites according to defined criteria (e.g. food product merchants only). The data acquisition module stores all matching data from webpages for later analysis and preservation of evidence. In another module, automatic recognition of a site's legal notice (impressum) is carried out for the respective vendor within whose online shop a potentially non-compliant food product is being offered. The image analysis tool performs logo recognition to enrich the text-based information of websites, thus providing additional visual information. The monitoring tool performs regular automated monitoring of e-food vendors, products and ingredients. The proof of principle of the prototype was achieved by conducting a web search for hazardous food products containing synephrine and caffeine. In total, 1242 product offerings on the internet for suspicious food products were identified among the 8683 search results. The software prototype has potential to enhance consumer protection and food safety with respect to e-foods.

1. Introduction

The birth of online shopping can be traced back to UK-resident Jane Snowball's purchase of groceries from Tesco over the internet in 1984 (Winterman & Kelly, 2013). Since then, both the selection of products on offer and revenues have increased considerably, with the market now representing a rapidly expanding sales channel with enormous growth potential. In 2012, food products worth €540 million were sold over the internet in Germany (Wagner & Wiehenbrauk, 2014). That translates to a growth rate of 400% over 2006 figures (BITKOM, 2007). UK residents are pioneers in online food shopping with 2012 internet sales totaling €5.5 billion, a 5% share of the overall market. The online food product market in Germany remains far below its full potential with a market share of only 0.3%. By 2020, however, a 33-fold increase to 10% market share has been forecast (Wagner & Wiehenbrauk, 2014).

All types of food and beverages can be purchased over the

[*] Corresponding author.
E-mail address: Jan.Fritsche@haw-hamburg.de (J. Fritsche).
[1] Seconded to the European Commission.

http://dx.doi.org/10.1016/j.foodcont.2015.09.039
0956-7135/© 2015 Elsevier Ltd. All rights reserved.

internet. There is an overwhelmingly large selection of goods spread across both small online shops and large websites which contain thousands of individual pages. These businesses include renowned enterprises that offer a full range of products such as Kaiser Tengelmann (since 1997), Edeka (entered the market in 2002), and Amazon (since 2010). Online marketplaces have recognized the trend, continually expanding their food product selections. EBay, for instance has 300,000 products listed in the gourmet category alone (eBay, 2015). The Chinese marketplace Alibaba.com, which also lists items for sale in the German language (german.alibaba.com), has over 40,000 producers and suppliers in the diet foods (Abnehmlebensmittel) category. Most of the products there are available only in commercial-size packages ranging from 1000 kg up to several tons (Alibaba, 2015). Furthermore the internet offers niche products and small merchants an ideal and modern platform which, considering the global nature of trade in the modern economy, is increasing in importance (Linder & Rennhak, 2012).

Not only well-known and adequately monitored goods are offered. Borderline products that blur the lines between food, cosmetics, and pharmaceuticals are also proliferating now (Löbell-Behrends et al., 2011, 2008), representing a potential health risk for consumers. These developments make it all the more important from the standpoint of health and consumer protection that the online trade in food products be recognized as an additional sales channel and be made an integral part of government controls.

The competent authorities are obligated in accordance with EC Regulation No. 882/2004 to undertake food controls based on a risk analysis with respect to health and fraud, regardless of which sales channel is being used. The European authorities however face the problem that stationary food inspections cannot be easily applied to internet sales (Schreiber, 2013). New, modified approaches are needed. Germany's Federal Office of Consumer Protection and Food Safety began working on a concept to monitor this new market segment as early as 2008. The objective was to achieve the same level of food safety online as currently found among bricks-and-mortar outlets (Köppe et al., 2014). An initial application of this concept is already being implemented for a real-world use case. Due to the search for high-risk products being limited to manual research, controls — although contributing significantly to online food safety — are currently unable to provide fully effective protection via economically feasible means. Monitoring the nearly limitless online marketplace for food products demands an automated, IT-based solution to implement efficient inspection procedures for this rapidly expanding, fast-paced market and ensure consumers are protected.

Below, we present a concept for automated e-commerce food product controls and provide a proof of principle based on a product research example.

2. Conceptual model

To develop the concept presented below, we will identify three relevant spheres of activity for automating online food product controls to the greatest extent possible: product controls, monitoring, and exposing the improper use of logos and seals. All three areas have in common that, until now, no customized automated solutions exist, the degree of attainable automation is estimated to be high, and an implementation is expected to increase the quality of the current manual monitoring of data used for food product controls.

The product controls as described in the present conceptual model seek to identify potentially non-compliant food products on the internet and report these to local authorities by determining the identifying information for the respective merchant. The automation will include both reactive and proactive controls. Reactive controls will serve to identify product offerings which have already been classified as high-risk. Proactive controls involve seeking out products and webpages which violate applicable food law but which the competent authorities have not yet become aware of. This approach is expressly called for in the corresponding literature (see Monakhova et al., 2011).

Relevant objectives which can be achieved through product controls range from identifying online product offerings with pharmacologically active ingredients (EC Regulation No. 178/2002 Art. 14) and detecting incorrectly or erroneously labeled products (EC Regulation No. 1169/2011) to identifying non-approved health claims (EC Regulation No. 1924/2006) and detecting products with non-approved novel foods (EC Regulation No. 258/97).

Monitoring food products traded over the internet should enable both proactive and reactive automatic controls on an ongoing basis in addition to risk assessment down to the level of specific merchants, products, and ingredients. This process should proceed in a very similar fashion to risk-based controls for conventional merchants. Vendors who repeatedly give reason for suspicion should be identified, as should emerging trends regarding non-compliant products. Follow-up monitoring after a predetermined interval should make it possible to detect violations that have not been corrected after a official measure has been taken and to uncover re-listed products which have been cited and removed in the past.

Detecting the improper use of logos and seals is an important part of the present concept for protecting consumers. In falsifying seals and logos, companies hope to achieve an economic advantage over other merchants by intentionally deceiving consumers. Especially with regard to organic products, high profit margins are a tempting incentive. Consequently, the present concept detects counterfeit seals and logos which do not correspond to the formal requirements of the seal issuers while also revealing products that bear seals and logos but do not have permission to use them. The main focus here is on the German organic seal and the EU organic logo.

2.1. Structure of the software's content

The functions depicted here are to be implemented in the form of a software prototype. The content of the software can be divided into three tools: the Research Tool, the Image Analysis Tool, and the Monitoring Tool. Each tool is itself divided into autonomous components. In each component within which the software reaches decisions based on its algorithms, it outputs a confidence value for the reliability of these decisions. This allows users to define thresholds. Decisions which are not deemed reliable enough can be entered later manually using a computer-aided evaluation interface. The prototype will for the most part be designed with a modular structure in order to ensure the software package remains flexible and permits specific computer-aided search tasks to be defined and carried out under consideration of configurable evaluation criteria.

2.1.1. Research tool

The research tool allows for the detection of potentially non-compliant products using internet search engines, the detection of violations of food regulations, and the compilation of the corresponding merchant information. In this tool, all information as text on the website is processed. The tool is divided into eight modules, which are displayed in Fig. 1.

The user is given the opportunity to set up custom research cases on various subjects of inquiry in order to receive — via internet search engine and marketplace search — as large a pool of webpages for further analysis as possible.

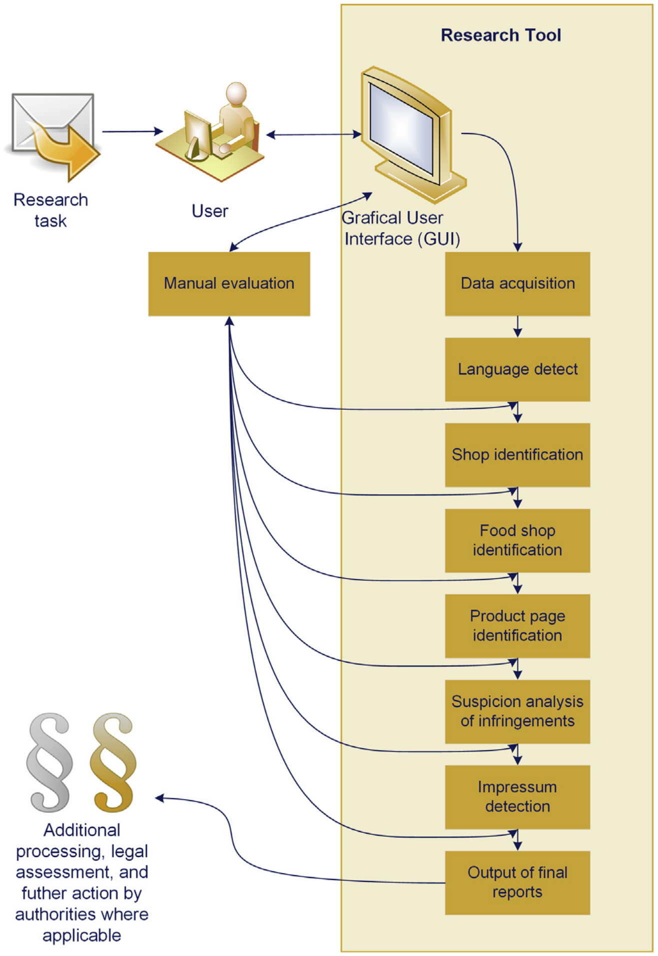
Research Tool
Research task
User
Grafical User Interface (GUI)
Manual evaluation
Data acquisition
Language detect
Shop identification
Food shop identification
Product page identification
Suspicion analysis of infringements
Impressum detection
Output of final reports
Additional processing, legal assessment, and futher action by authorities where applicable

Fig. 1. Research Tool process for automated controls of food products sold via electronic commerce.

After the desired search terms have been specified, the automated analysis process begins. First, a data acquisition module stores all matching data from webpages meeting the defined criteria and saves local copies of the pages to the software's dedicated database for later analysis and preservation of evidence. These data include the matching webpages with URL, their respective home page, and the corresponding impressum (site legal notice). Duplicates results from the same search engine or from different search engines are automatically detected during the data acquisition process and only considered once from then on.

A multi-stage filtering process is performed to limit the sites first to German-language results, then to shop sites, food product merchants only, and then according to the respective product offering pages. Within the suspicion analysis of infringements, these offering pages are automated by means of text analysis using criteria that can point to food safety violations, with all corresponding information being stored in the database. Both proactive and reactive procedures are used here. In other words, each page is first analyzed with regard to the presence of a previously identified violation. In addition, all other specified food safety regulatory criteria (which are automated or can be verified by means of websites) are checked on the site. In this manner, proactive controls are carried out which can uncover additional potential violations that were not previously known.

In another module, automatic detection of the legal notice (Impressum) is performed for the respective vendor within whose online shop a potentially non-compliant food product (for instance with hazardous ingredients) is being offered or for whom a different violation against food law has been determined (missing ingredients list, for instance).

In addition, the prototype automatically matches websites with potential violations to a specific EU country, non-EU country, or district in Germany based on data from the legal notice on the site. All of a merchant's relevant information is automatically summarized in a final report which is transmitted to the competent local authorities, providing the basis for further official action.

2.1.2. *Image analysis tool*

The functionality of the image recognition tool consists of recognizing, storing, analyzing, and evaluating all questions in conjunction with processing information present as images. Text needs to be extracted from images and made available for investigation or legal site notice recognition, similar to the way the research tool functions. Logos and seals should also be recognized and identified by comparing them to lists and images in order to detect violations.

2.1.3. *Monitoring tool*

The monitoring tool performs repeated automated monitoring of food product e-commerce. Specifically, it automatically triggers targeted follow-up monitoring for previously cited online product offerings. This is carried out in different phases:

First, it is determined whether or not the cited products are still online. This is an enforcement instrument comparable to follow-up inspections by a food inspector who checks to ensure that previously cited violations have been corrected.

Then, risk-based (in the broadest sense of the term) follow-up monitoring for the previously conducted research is performed. This means that an existing case that has already been handled triggers an additional investigation at a later time. The time-based weighting and order of the investigations conducted is automatically determined according to parameters with different weightings, such as the number of matches or priority level.

To implement the proactive monitoring, suspicious vendors (ones that have in the past been frequently cited) can be sought out across the entire internet site including all sub-pages and web shops with the assistance of a web crawler to perform additional analyses of food law violations.

2.2. *Software architecture and technologies*

As illustrated in the conceptual model described above, a software application is needed for automated food safety monitoring on the World Wide Web that can process websites and filter them successively using analysis tools and according to specified criteria in addition to extracting content. For the current software prototype, an architecture and technologies are required that can meet these requirements. Below, both the software architecture with the defined components and the corresponding technologies are explained.

2.2.1. *Software architecture*

On the technological level, the software consists of subcomponents which together provide the full functionality. These subcomponents are selected and built according to what are known as architectural patterns and design patterns. The selected patterns have different advantages, for instance for (ongoing) development and software recyclability. In this manner, it's possible to replace components and technologies if required and adapt them to future needs (Fowler, 2002).

For these requirements to be fulfilled, a three-layer architecture was chosen for the software (see Fig. 2), which makes it possible to easily swap out technologies (Eckerson, 1995; Jing-Feng, Li, & Chen, 2006). Consequently, it would be possible to replace the front end with a different web application or add different components to the back end to supply additional functionality.

2.2.1.1. Client tier. A web application serves as the prototype's presentation layer. The selected technologies on the presentation layer comprise a combination of JavaScript libraries with HTML and CSS. This combination has emerged as the de facto standard for modern web application development. The main component of the client tier is the graphical user interface (GUI), which is responsible for displaying the results from the application layer and passing on user input to the application tier. This layer forms the front end of the entire application.

2.2.1.2. Application tier. On the application layer, the Web Application processes inputs, collects search engine results using the scraper, crawls domain content, and controls the Analyzing Application. A scraper is an application that extracts content from documents using a number of different technologies. Within the software prototype, the technology identifies the search results from search engines and stores them. Crawlers, on the other hand, are used to read hyperlinks on certain websites by automatically following their link structures. This is intended to make it possible to include specific web pages (e.g. the page with the legal contact information of a merchant) or all individual pages of a website. The Analyzing Application is based on C#, which includes suitable libraries and tools for data analysis. In the Analyzing Application, all data are processed and analyzed for specified conditions. For instance, using what are known as regular expressions, addresses on webpages are found and an automatic evaluation is attempted to determine whether or not the page in question is a shop. If no automatic evaluation can be carried out or the results are uncertain, a manual follow-up is triggered. Since an all-encompassing specification for what constitutes a web shop would be very complex for an algorithm, the system learns the definition through user input and manual follow-up evaluations. As a result of this continual expansion of the available pool of data, the success rate keeps improving at each filter level. To ensure that as much data (websites) as possible can be processed by the automated analysis, this module has been designed as a scalable, parallel computing application. The software is multithreading-capable on a single computer to utilize multiple cores (threads) simultaneously but also can run in parallel on multiple networked computers (both Windows and Linux).

The components for implementing the filter levels, referred to below as modules, have special significance for the Analyzing Application on the back end. Interfaces must be defined for the individual modules so that their functionality can be integrated into the overall software. Consequently, they work in a similar fashion to add-ons, which add additional functionality to existing software.

There is an additional special module which does not have a predetermined function, but rather can be used as a multipurpose tool to process and link data. This module can be used to create custom filters. It handles incoming data mainly in the form of tables. These tables can for instance be processed and filtered using SQL commands.

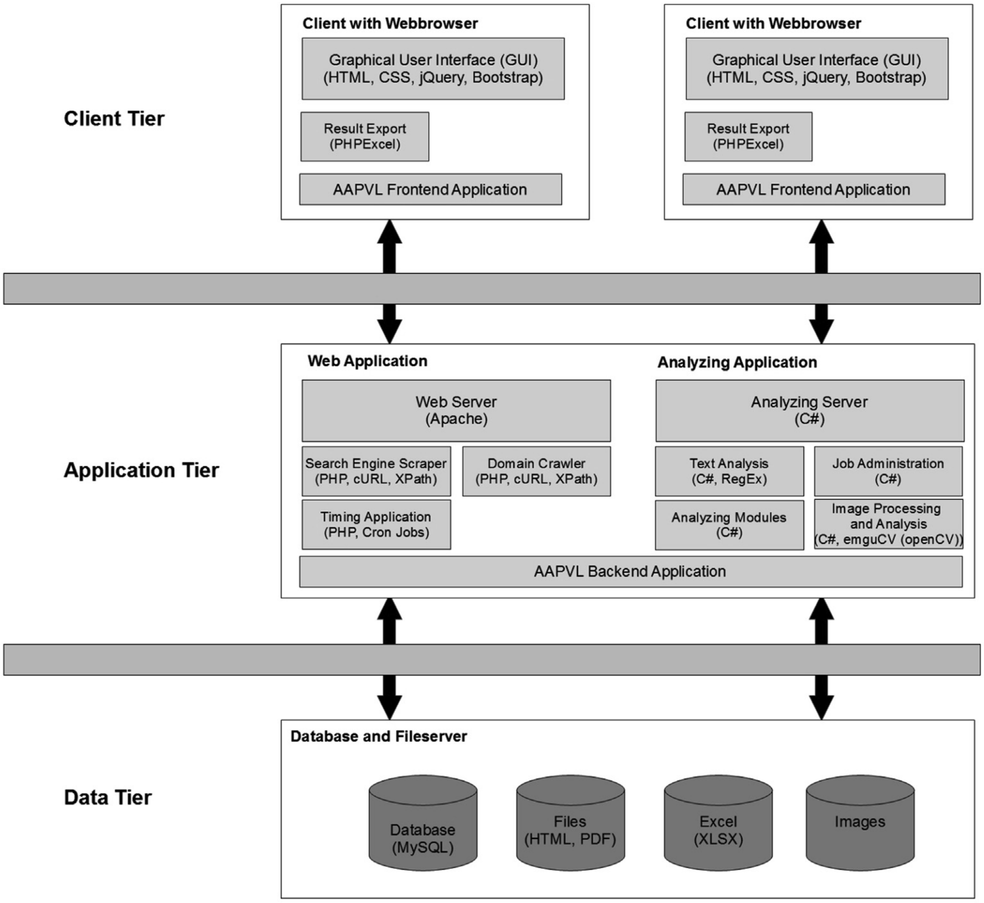
Client Tier
Client with Webbrowser
Graphical User Interface (GUI) (HTML, CSS, jQuery, Bootstrap)
Result Export (PHPExcel)
AAPVL Frontend Application
Client with Webbrowser
Graphical User Interface (GUI) (HTML, CSS, jQuery, Bootstrap)
Result Export (PHPExcel)
AAPVL Frontend Application
Application Tier
Web Application
Web Server (Apache)
Search Engine Scraper (PHP, cURL, XPath)
Domain Crawler (PHP, cURL, XPath)
Timing Application (PHP, Cron Jobs)
Analyzing Application
Analyzing Server (C#)
Text Analysis (C#, RegEx)
Job Administration (C#)
Analyzing Modules (C#)
Image Processing and Analysis (C#, emguCV (openCV))
AAPVL Backend Application
Data Tier
Database and Fileserver
Database (MySQL)
Files (HTML, PDF)
Excel (XLSX)
Images

Fig. 2. Software architecture of the prototype.

2.2.2. Data storage tier

Applications on the application tier communicate by means of a MySQL database on the data tier. All results from data acquisition (stored websites, stored images) are stored on the data layer. Metadata for the content is stored in the MySQL database. The Web Application and Analyzing Application use this shared database as an interface to communicate with one another. Both applications write and read data required for individual tasks here.

The Analyzing Application thus receives commands as well as the data content to be analyzed from the Web Application. The front end on the presentation layer (client tier) is used to display the analysis results. To accomplish this, it receives the data to be displayed from the Web Application which accesses the shared MySQL database.

2.2.3. Software components on the layers

As mentioned above, the software consists of various components which, together, deliver the complete functionality. Table 1 lists the components the software comprises and which specific technologies are used to form the basic components.

3. Proof of principle

The product control concept was tested using an example case in collaboration with the Bavarian Health and Food Safety Authority (Bayerischen Landesamt für Gesundheit und Lebensmittelsicherheit). As part of extensive online investigation, product offerings were researched that — according to their promotional content and packaging — contained synephrine and

Table 1
Main technologies used for the software components on the three layers.

Components	Technologies	Description
Client tier:		
Graphical User Interface (GUI)	HTML, CSS, jQuery, Bootstrap	The graphical user interface has been implemented in the form of a web interface that can be accessed using any standard web browser (e.g. Internet Explorer, Mozilla Firefox). The technologies used for the GUI include a combination of HTML (hypertext markup language) and a text-based markup language for structured representations of digital documents. CSS (Cascading Style Sheets), a stylesheet language used to format content. HTML and CSS are web standards managed by the W3C. Whereas HTML determines the content structure, CSS specifies the style. This results in an intended separation of content and layout. jQuery and Bootstrap are libraries based on JavaScript that are used for instance to process forms, represent design elements, and conduct asynchronous data processing. JavaScript is a scripting language developed for DHTML (dynamic HTML). JavaScript makes it possible to reload content on websites in real-time, for example. On the modern web, JavaScript also plays a very important role in developing apps for smartphones.
Result export	PHPExcel	PHPExcel is a PHP library that is free to use and serves as an interface between spreadsheet programs and PHP. PHPExcel provides a number of functions which are helpful in reading and generating Excel documents. For the software prototype, results from the analyses are converted into spreadsheets, which allows them to be viewed for instance in Microsoft Excel.
Application tier:		
Web server	Apache	The Apache web server is an open source product that sees the most use globally for hosting HTTP servers. Apache has a modular structure and includes a number of modules that can be used as a basis for web applications. In the present software prototype, mainly the PHP modules and cURL library (client for URLs) are used.
Search Engine Scraper	PHP, cURL, XPath	The Search Engine Scraper is an internally developed application for automated querying of search engines. The search results are then stored by the Scraper and made accessible to the Analyzing Application. The scraper is based on the cURL application, which imitates the behavior of a web browser. This allows it to read and store the contents of websites. With the help of XML Path Language (XPath), structured information can be read from XML documents and HTML documents as well, since HTML is also used to structure digital documents. For the prototype, information defined via XPath (e.g. URLs of search results and search engine results pages, aka SERPS) is extracted.
Domain crawler	PHP, cURL, XPath	The Domain Crawler application is also an in-house development and is based on the same technologies as the Search Engine Scraper. It just has a different configuration. The Domain Crawler is used to find and store additional URLs on websites. This makes it possible for instance to read all websites on a specific domain and make them available for analysis.
Timing Application	PHP, cron jobs	The Timing Application is mainly an extension of the functionality for defining search queries used by the Search Engine Scraper. The additional information submitted regarding the time of the scraping or specifying submission of search queries at regular intervals allows multiple analysis processes to run in the background. The Timing Application is a critical component of the monitoring process. It is implemented using what are known as cron jobs defined within a cron daemon. Cron daemons are used in Linux environments to automatically perform recurring tasks.
Job administration	C#, Locks, Threads, MySQL	To allow distributed operation across the network, a job scheduler has been included that independently distributes the tasks to be completed to the individual units (workers) which then process the tasks. Any unit can automatically assume the duties of the job scheduler at any time if the respective task has not already be checked out by another instance on the network.
Analyzing modules	C#, RegEx	The individual modules for implementing the filters are designed as add-ons. Modules can work independently or receive and exchange data via specified interfaces.
Image Processing and Analysis	C#, emguCV	For the analysis and evaluation of image data, such as for instance recognizing organic logos, an image recognition component based on the emguCV (C# openCV wrapper) library is planned.
Data storage tier:		
Database	MySQL	The database for exchanging information between the front end and back end applications has been implemented using MySQL. MySQL is one of the most widely used open-source database management systems (DBMS) for relational databases.
Files	HTML, PDF	Text-based documents that are used as raw data for the Analyzing Application. The documents for text analysis are available as HTML documents (websites) or in the form of PDF documents.
Excel	XLS, XSLX	The analysis results are exported to spreadsheets. Using PHPExcel, XLS or XLSX documents are generated. Spreadsheets can also be imported and used as an initial data source. Such as for instance a list of URLs that need to be processed.
Images	JPG, PNG, BMP, GIF, TIFF	Stored images in all conceivable formats for image analysis.

caffeine. Synephrine, particularly in combination with caffeine, is suspected to be hazardous to health (Bundesinstitut für Risikobewertung, 2012; Firenzuoli, Gori, & Galapai, 2005; Rossato, Costa, Limberger, Bastos, & Remiao, 2011; Stephensen & Sarlay, 2009). The research started with 11 specific search queries which were all submitted to both google.de and bing.de. Data was collected using the data acquisition model described above. Specifically, a scraper was used to automatically send the search queries to the search engines, collect the results, and deduplicate.

In addition, the results were manually screened for relevance. This procedure corresponded for the most part to the multilevel filtering of the Research Tool. The Evaluation Tool that serves to enable manual follow-ups was used to support the evaluation of the search results.

3.1. *Results from the proof of principle*

The investigation conducted via search engines resulted in 8683 results. Approximately 63% of these results were identical URLs (duplicates) either listed twice in the same search engine or returned by both search engines (Fig. 3). An additional 4% of the search engine results were unreachable websites. After this automated screening, 2935 results were provided for relevance evaluation.

In approximately 42% of these cases (1242 results), the results were relevant product offerings (Fig. 3). Over half (57%) of the search engine results represented irrelevant pages according to the relevance evaluation. This included both pages which did not have any product offerings as well as pages with product offerings that did not correspond to the target of the investigation. Pages that could not be clearly categorized were marked as inconclusive search results (19 results), including for instance shops that were temporarily unreachable due to maintenance. Only seven search engine results were in a language other than German. The majority (63%) of the 1242 identified relevant offerings were from merchants located in Germany (Fig. 4).

Of the 1242 relevant search engine results, a total of 219 different products were found that match the investigation target. These products were being sold in 449 web shops. The number of web shops is calculated based on the number of different domains of the results webpages. Within Germany, the shops in which the products are offered are distributed geographically across the entire country, except for Saarland (Fig. 5).

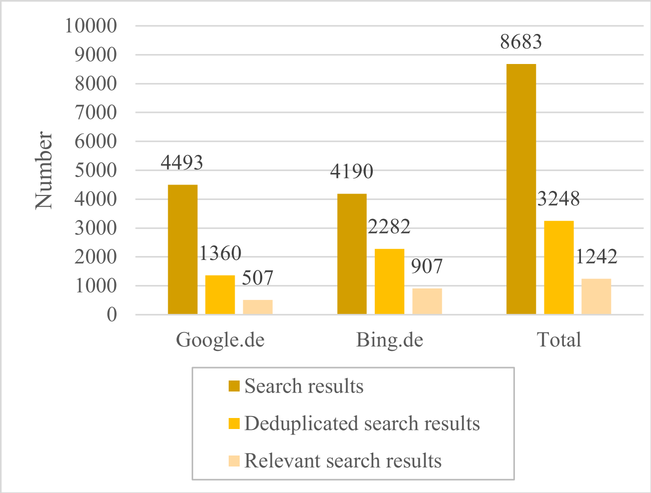

Fig. 3. Number of search results and duplicates per search engine and overview for 11 search queries.

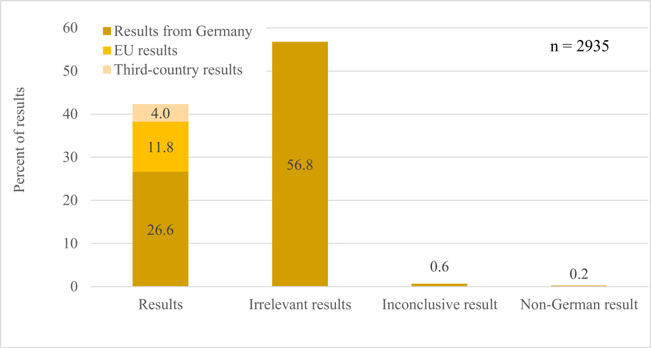

Fig. 4. Distribution of the search results from 11 search queries submitted to google.de and bing.de for results categories and distribution of the geographical origin of the respective hits.

The 1667 irrelevant search results represent the entire spectrum of different types of websites (Fig. 6). The most frequently represented ones in the unfiltered data are blogs and forums in which people exchange information about the respective products and substances. But shops selling food products that did not offer any of the products being searched for were also included by the search engine. Often, this was a result of the website emphasizing that the respective substance was not included in their products or that the offered products have a similar effect to the targeted substances. In some cases, the offering was categorized as irrelevant since only one of the targeted substances was included according to the ingredients list.

4. Discussion

The current research makes apparent the vast scope involved in product controls on the internet. The quantities of data generated are exceedingly difficult to process manually. It has been demonstrated that the application of the current concept for product controls with its multi-level filters exhibits high efficacy. The number of results pages found through search engines was able to be systematically reduced to only those with relevance. In this manner, it was possible to present the Bavarian Health and Food Safety Authority with a pre-processed package of relevant information.

Initiating the product control process with search engines is for the most part identical to the approach used by Kuhr, Krewinkel, Raschke, and Schreiber (2012). The concept thus clearly emphasizes proactive consumer protection as opposed to general monitoring of all internet commerce. Consequently, those product offerings should be monitored that a consumer with an average knowledge level will find in online shops. Since Google and Bing together represent a 92% share of the search engine market (Webhits 2014), this prominent position makes it essential for merchants to be listed in these search engines if they wish to be found by prospective customers. This situation is exploited in the current concept by using search engines to investigate.

In this concept, a decision was made not to develop a specialized web crawler that would independently search the internet and build its own dedicated web index, since billions of URLs would need to be processed, of which the majority would be irrelevant for the subsequent analysis (Patterson, 2004). The benefits would be far outweighed by the considerable additional financial and development resources required (see Patterson, 2004). The exact proportion of the entire World Wide Web that is indexed by search engines has not been sufficiently researched, but the assumption is

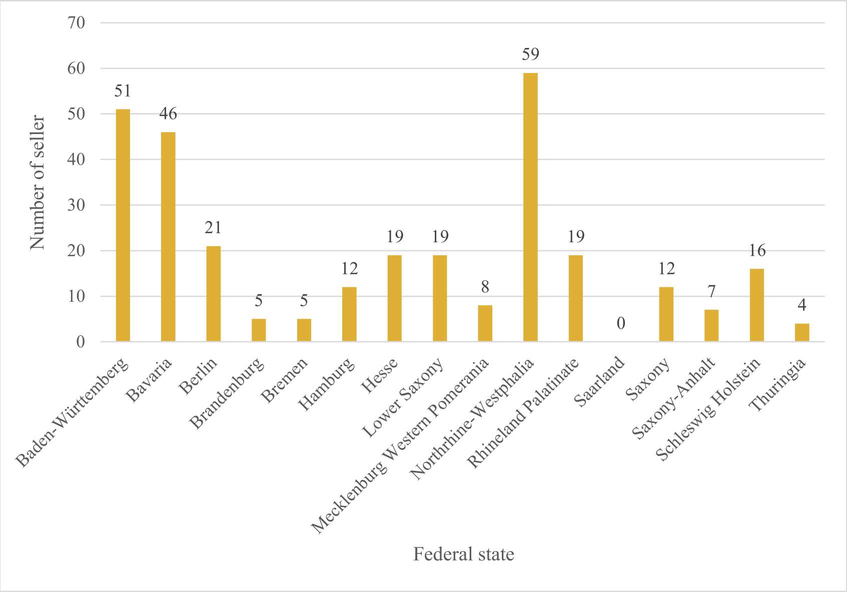

Fig. 5. Number of merchants of corresponding product offerings per federal state in Germany.

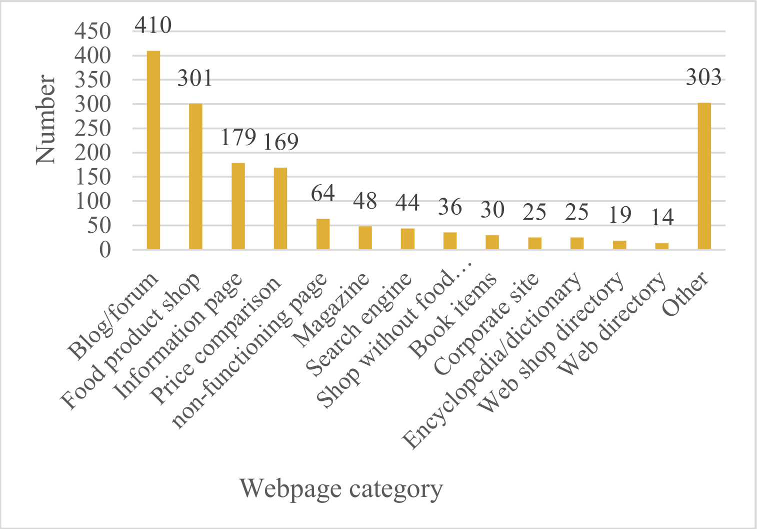

Fig. 6. Distribution of the irrelevant results for the search engine query on the different website types.

that the portion of the World Wide Web used by consumers is limited to those sites which are indexed by search engines, the so called Surface Web (Lewandowski & Höchstötter, 2007). It is however undisputed that the number of indexed pages differs from one search engine to the next and thus indexing more than one search engine is an appropriate means of ensuring greater coverage, even if this does mean collecting a large number of duplicates (Bharat & Broder, 1998; Gulli & Signorini, 2005).

In contrast to the approach of Kuhr et al. (2012), search queries were not used as the first filter (see Krewinkel, Tolg, & Fritsche, 2011). Instead, the search engines were employed in the opposite manner as a tool used specifically to generate a universe (pool of data) that is subsequently screened using sequences of configurable filters until only relevant results are left. Since the concept provides for a largely automated execution of the sequential filters which follow, it is not necessary to limit the quantity of data beforehand.

Using configurable filters has the advantage of providing known and modifiable parameters. Imposing constraints by means of the search engines themselves however proves more difficult. In the relevant academic literature on search engines, there are indications that difficulties exist with the quantity of results for complex search requests (see discussion in Jacsó, 2006; Notess, 2003). Uyar (2009) demonstrated that, as the number of elements in the search term increased, the precision of results decreased.

Furthermore, analyses of the behavior of search engine users show that a search with complex search terms would differ greatly from the typical search behavior of an average user. Höchstötter and Koch (2009) showed in a comprehensive analysis of search requests that the search query length averages 1.8 words. More complex requests represented only less than 3% of queries (Schmidt-Maenz & Koch, 2006). Only half of search engine users are familiar with Boolean operators. And those who use them frequently do so incorrectly (Jansen, Spink, & Saracevic, 2000; Machill, 2003).

After selecting the universe, selecting filters is a very important step for the concept. The effectiveness of filters with respect to increasing the degree of relevance of the search results is exhibited by the evaluation of the duplicates and the categories of the irrelevant results. 63% of the results were already eliminated by the first of the filters, deduplication. In the current testing, this was the most effective filter. Manual deduplication is not practicable with over 5000 entries. The second filter, which differentiated between shops and non-shops, proved to also be exceptionally effective, achieving a reduction rate of 35%. Less clear, initially, is the efficacy of the food product shop detection. The reduction attained in the current example amounted to just 2%. Prior investigations have however shown that this number can vary greatly from case to case, since — depending on the targeted products or substances — similar names can be expected to appear in non-food sectors which have their own shops for online sales. In such cases in particular, the screening filter and also the product page identification — the effectiveness of which was not specifically examined in the current study — are obligatory for automating the process of investigating suspected violations. Only once it has been determined that the respective website is a web shop offering food products and that the web page represents a product offering on that site, can the prototype analyze the page for potential violations.

Using the example case, it becomes clear that the language detection filter did not result in the desired effect: an additional reduction of the relevant search results. Only 7 of the 2935 results were non-German-language pages. From this we can conclude that the language selection in the advanced settings of the search engine works quite reliably when German is selected as the language on the localized German-language search engine site (Lewandowski, 2008). As such, an additional verification using this filter is unnecessary.

For real-world electronic commerce food product controls, verifying company registrations and product sampling remain important components of an overall product monitoring concept.

5. Conclusions

Based on the current example case, the feasibility and practicality of the concept as implemented for online food product controls were demonstrated. Additional test scenarios must now be completed in order to acquire the necessary knowledge about the structures and changes taking place in the fast-paced world of online commerce. Of particular significance here are the criteria for investigating, monitoring, and uncovering fraud through the misuse of logos or seals. An additional focus on artificial intelligence is also conceivable. Furthermore, the criteria mentioned above must be implemented in software prototypes once they have been developed, with the reliability and validity of the filters used in these prototypes verified through additional test scenarios. Overall, the concept presented here for monitoring internet-based food product commerce is able in an efficient manner to make an important contribution towards achieving a degree of food product safety on the internet that equals what is established for the bricks and mortar merchants.

Acknowledgment

The project is supported by funds of the Federal Ministry of Food and Agriculture (BMEL) 2819102112 based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme. The project is carried out in cooperation with the German Federal Office of Consumer protection and Food Safety (BVL).

References

Alibaba Group. (2015). *Category overview, Abnehmenlebensmittel*. http://german.alibaba.com/Slimming-Food_pid100006931 [Retrieved Jun 08 2015].

Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM). (2007). *Press release: Deutsche bestellen Lebensmittel für 122 Millionen Euro im Internet*.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems, 30*, 379–388.

Bundesinstitut für Risikobewertung. (2012). *Gesundheitliche Bewertung von synephrin- und koffeinhaltigen Sportlerprodukten und Schlankheitsmitteln. Stellungnahme Nr. 004/2013 des BfR from November 16, 2012*.

eBay Inc. (2015). *Kategorieübersicht Feinschmecker*. http://www.ebay.de/sch/i.html?_from=R40&_trksid=p2050601.m570.l1313.TR0.TRC0.H0.TRS0&_nkw=&_sacat=62682 [Retrieved Jun 08 2015].

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

Firenzuoli, F., Gori, L., & Galapai, C. (2005). Adverse reaction to an adrenergic herbal extract (Cit-rus aurantium). *Phytomedicine, 12*(3), 247–248.

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

Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *Proceedings of the 14th international conferece on World Wide Web. Chiba, Japan*.

Höchstötter, N., & Koch, M. (2009). Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science, 35*, 45–65.

Jacsó, P. (2006). Dubious hit counts and cuckoo's eggs. *Online Information Review, 30*(2), 188–193.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management, 36*, 207–227.

Jing-Feng, L., Li, Y., & Chen, P. (2006). A unified architecture model of web applications. *Journal of Shanghai University, 6*(3), 221–227.

Köppe, D., Harms, H., Kranz, P., Krewinkel, A., Kuhr, C., Rachke, D., et al. (2014). Die Gemeinsame Projektzentralstelle „Kontrolle der im Internet gehandelten Erzeugnisse des LFGB und Tabakerzeugnisse“ (G@ZIELT) – Vorstellung und erste Ergebnisse. *Der Lebensmittelkontrolleur, 2*, 9–12.

Krewinkel, A., Tolg, B., & Fritsche, J. (2011). Online-Lebensmittelhandel und Strategien zur Kontrolle des virtuellen Lebensmittelmarktes. *Journal of Consumer Protection and Food Safety, 6*, 395–400.

Kuhr, C., Krewinkel, A., Raschke, D., & Schreiber, G. A. (2012). Der Internethandel – erste Ergebnisse der Zentralstelle im BVL. *Rundschau für Fleischhygiene und Lebensmittelüberwachung, 64*, 247–249.

Lewandowski, D. (2008). Problems with the use of web search engines to find results in foreign languages. *Online Information Review, 32*(5), 668–672.

Lewandowski, D., & Höchstötter, N. (2007). Web searching: a quality measurement perspective. In M. Zimmer, & A. Spink (Eds.), *Web search: Interdisciplinary perspectives*. Dordrecht: Springer.

Linder, M., & Rennhak, C. (2012). *Lebensmittel-onlinehandel in Deutschland. Reutlingen working papers on marketing & management*. ESB Business School, Hochschule Reutlingen. Nr. 2012-4.

Löbell-Behrends, S., Böse, W., Maixner, S., Kratz, E., Kohl-Himmelseher, M., Marx, G., et al. (2011). Kontrolle des Internethandels mit Lebensmitteln: Abgrenzung bei Borderline-Produkten und Ansätze für effektive Kontrollstrategien. *Journal of Consumer Protection and Food Safety, 6*, 385–393.

Löbell-Behrends, S., Maixner, S., Kratz, E., Kohl-Himmelseher, M., Bauer-Aymanns, H., Marx, G., et al. (2008). BORDERLINEPRODUKTE Kontrolle des Internethandels mit Anti-Aging- und Schlankheitsmitteln. Eine Pilot-Studie. *Deutsche Lebensmittelrundschau, 104*(6), 265–270.

Machill, M. (2003). *Wegweiser im Netz Qualität und Nutzung von Suchmaschinen*. Gütersloh: Verlag Bertelsmann Stiftung.

Monakhova, Y. B., Löbell-Behrends, S., Maixner, S., Böse, W., Marx, G., & Lachenmeier, D. W. (2011). Automated classification of web pages for identification of suspicious food products – a feasibility study. *Rundschau für Fleischhygiene und Lebensmittelüberwachung, 107*, 328–330.

Notess, G. R. (2003). *Google inconsistencies*. http://www.searchengineshowdown.com/features/google/inconsistent.shtml [Retrieved Dec 15 2014].

Patterson, A. (2004). Why writing your own search engine is hard. *Queue - Search Engines, 2*(2), 48–53. April 2004.

Rossato, L. G., Costa, V. M., Limberger, R. P., Bastos, M. L., & Remiao, F. (2011). Synephrine: from trace concentrations to massive consumption in weight-loss. *Food and Chemical Toxicology, 49*(1), 8–16.

Schmidt-Maenz, N., & Koch, M. (2006). A general classification of (Search) queries and terms. In *Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)*.

Schreiber, G. A. (2013). Lebensmittelkontrolle und Verbraucherinformation im digitalen Zeitalter. *Journal of Consumer Protection and Food Safety, 8*, 267–269.

Stephensen, C. T. A., & Sarlay, M. R. (2009). Ventricular fibrillation associated with use of syneph-rine containing dietary supplement. *Military Medicine, 174*(12), 1313–1319.

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*, 469–480.

Wagner, W., & Wiehenbrauk, D. (2014). *Study: Cross Channel Revolution im Lebensmittelhandel. Ernst & Young GmbH*.

Webhits.de. (2014). *Suchmaschinen*. http://www.webhits.de/artwork/ws_engines_druck.png [Retrieved Dec 15 2014].

Winterman, D., & Kelly, J. (2013). *Online shopping: The pensioner who pioneered a home shopping revolution. BBC news magazine*. http://www.bbc.com/news/magazine-24091393 [Retrieved Jan 16 2015].