Development of a food safety information database for Greater China

Shanquan Chen [1], Dandan Huang, Wenyan Nong, Hoi Shan Kwan*

The School of Life Sciences, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China

A B S T R A C T

Food safety in China has become a major research interest and popular social issue in recent years. We have constructed a database-website-advanced search engine system for continuous collection online food safety information and efficient mass literature analysis. This paper will introduce the methods used for data extraction, processing, search and intelligent result analyses. Over 1,300,000 pieces of entry have been collected, including the information from the websites of government agencies, the abstracts of academic papers and news in credible online media. This system can catalog the search results automatically in terms of geographic region, time period, original sources and food-related keywords. An application is shown through comparing the differences between food-borne disease data in governmental documents and media reports. The number of food poisoning incidents reported by the China Center for Disease Control and Prevention (CDC) decreased gradually from 2006 to 2012 whereas the number of food poisoning news items increased simultaneously with the food safety news. The media prefer to report the food safety incidents caused by man-made hazards or those which occur in the public domain.

1. Introduction

During the past decade, the occurrence of food safety issues, such as adulteration, additive violation and microbiological contamination, has become more frequent and the influence range has increased exponentially with the globalization of food chains. China has been in the center of this vortex for many years because of its recurrent scandals. The notorious incidents include melamine-adulterated milk products in 2008 which led to the death of six infants, reproduced stale meat from the Chinese unit of the OSI Group which serves many international fast-food chains, and ditch oil; these scandals frequently make headline news. With the increasing volume of food exports, Chinese food safety issues have not just impaired the health of the domestic population but also caused deep concern around the world. Scientists, officials, industries, teachers and the public are all demanding comprehensive food safety information of Greater China.

Unlike the environmental toxins or unintended infectious agents in developed countries, China's food safety contaminations are mainly caused by illegal chemical additives introduced intentionally (The Lancet, 2014). In order to analyze them and ensure prevention, it is critical to review the information about previous incidents, known problematic ingredients and detection methods. Van de Brug, Lucas Luijckx, Cnossen, and Houben (2014) analyzed scientific papers, the documents released by food safety authorities and related news to collect the information on early signals preceding the emergency of former food safety incidents. Moore, Spink, and Lipp (2012) collected the scientific papers and English articles in the general media about food fraud and adulteration in order to develop a database and to analyze the data for trend identification. However, there has hitherto been no systematic effort to compile the information on food-safety-related knowledge available in Chinese-language news media, government sites and academic articles.

Academic papers are a precious pool of technical information for indepth analysis and comparison. In Mainland China, the majority of scientific papers are collected into the China Knowledge Resource Integrated Database (http://www.global.cnki.net/kns50/) and can be searched through the service platform of the China National Knowledge Infrastructure (CNKI). The data from these authorities are relatively credible and practical for supervision purposes. The Chinese governmental food safety information

* Corresponding author. Rm 288, Science Centre, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China.
E-mail addresses: chenshanquan@gmail.com (S. Chen), mikey.huang2011@gmail.com (D. Huang), nongwy@gmail.com (W. Nong), hskwan288@gmail.com, hoishankwan@cuhk.edu.hk, hskwan@eservices.cuhk.edu.hk (H.S. Kwan).
[1] Present address: The School of Humanities & Social Science, The Chinese University of Hong Kong (Shenzhen), Longgang District, Shenzhen, PR China.

system is very complex, and the data from the food safety authorities are scattered around the official websites of different departments. It is necessary to establish a one-step platform to integrate multi-source information from this complex system. Meanwhile, food safety news is an important source of timely information. According to the WHO, unofficial informal sources contribute more than 60% of the initial outbreak reports (WHO, 2015). Some scientists also utilized food safety reports on the internet to understand the nature of the food safety incidents which occur in China. Qiang, Wen, Jing, and Yue (2011) downloaded 600 reports dating from 1st April to 30th June, 2009 to investigate different subjects by content analysis (Qiang et al., 2011). In 2015, four other Chinese scientists analyzed 295 food safety incident reports, which occurred in Beijing, in an internet database from 2004 to 2013 (Liu, Liu, Zhang, & Gao, 2015). Such kinds of paper and tools require a large volume of relevant data. But most databases which focus on food safety news in Mainland China are edited by people without a professional background and who have not opened up their source list. These might impose an impartiality risk to the collected data.

As a result, our laboratory established a Food Safety Information Database to collect food safety news from the websites of the listed media, information from the official websites of the governmental agencies and related abstracts from Chinese-language journals in Greater China. To collect the data continuously, the procedures are automated with the assistance of information technology. We also developed some statistical functions in the user-interface, a website, to make this system a semantic tool of information searching and analysis. This paper will introduce the method of building this database-based web system, list the source websites and demonstrate how to use them. Then, we will compare the food-poisoning data in governmental documents and in media reports to illustrate a new application.

2. Materials and methodology

Initially, the scope of the target information was defined by a list of food safety websites in Greater China. The technique of web-crawler and Solr, open source software, were used to extract the raw data and index them after processing. In order to search and classify the collected information, a comprehensive list of food-safety-related keywords were consolidated from three glossaries in different fields. The number of keywords which appeared in each record was counted and stored in the database with the content. Furthermore, a website was built by using Java-script as the user interface, and a serial of plug-ins were developed based on the results of the keyword statistics. Content analysis can be conducted on the collected data to characterize the pattern of published food safety information by calculating the keyword frequency and generating the tendency profile automatically.

2.1. Source identification

The relevancy of collected information is guaranteed by the source websites. The food safety expert in our laboratory defined 106 websites as the information source of this Food Safety Database. They consist of three groups: the websites of governmental agencies in Greater China; the Chinese literature platform CNKI; and the media. The official food safety information system in China is complex. Initially, the food safety supervision system included the State Council's Food Safety Committee, the Ministry of Health (MOH), the State Food and Drug Administration Department (SFDA), the Ministry of Agriculture (MOA), the Industry and Commerce Department (IAC) and the Administration of Quality Supervision, Inspection and Quarantine Department (AQSIQ). In March

2013, reformation and functions among these departments was re-organized to simplify the system and to define their responsibility. The departments involved were consolidated into the China Food and Drug Administration (CFDA), the National Health and Family Planning Commission (NHFPC), MOA, AQSIQ and the State Administration for Industry and Commerce Department (SAIC). Moreover, the state and local food safety networks were merged into the food safety information system. The local level consisted of a hierarchy of three tiers of governments: provinces; cities; and counties (Jia & Jukes, 2013). The state administrations publish information with national impact, and the local administrative agencies announce the information within their geographic jurisdictions. China has 34 provincial administrative regions, but some of them do not have official websites, and the situation is even worse at the city level. The governmental agency group contains 54 websites both at the central and the local level, including the official websites of the supervision bodies, the affiliate websites for information communication and the websites of related institutions. There is one source in the academic group, CNKI, for the collection of academic abstracts. The rest of the related websites are categorized into the media group. We considered three kinds of websites as credible and/or valuable. The first is the websites written or edited by people with a food science or food safety background, including the official websites of the food research institutions and bloggers of famous food scientists. The second is the websites for the industry practitioners, such as the websites of the industry associations and the related companies. The third is the food or health columns of the portals and mainstream media which cover the news read by most people. Table A lists the source websites of this database-website system.

2.2. Data extraction, processing and storage

We deployed the technique of web-crawler to extract all information from the source website for information completeness. The raw data are duplicated from the identified data sources and the characters of each record are identified, including the title, source, publication date and URL. The content and characters of the processed data are stored in a database established with MySQL. The open-source software Solr can provide a full-text search and a replicate index. It imports records from the database, segments them into words by MMSeg4j, filters out the stop words and indexes the rest of the segmented words.

The extracted information contains some technical terms, so we constructed an inhouse glossary related to food science and safety to make the segmentation more precise. This inhouse glossary has 51,621 keywords. It contains the words from "A Glossary of Food Science and Technology", "Glossary of Biotechnology for Food and Agriculture" and the keywords in "throwing out the window net". Chau and Kwan (2005) compiled "A glossary of Food Science and Technology" which covers over 20,000 food-related terms with their Chinese translation used in Mainland, Taiwan and Hong Kong. These terms are from the areas of raw materials, processing, food hygiene, micro-organisms, nutrition, chemistry, food science analysis and additives. Food and Agriculture Organization of the United Nations (FAO) consolidated the terms and acronyms that are regularly used in the biotechnology for food and agriculture in "Glossary of Biotechnology for Food and Agriculture" (FAO, Ma & Zhao, 2011). We added the second edition with its Chinese translations both in simplified and traditional Chinese. Moreover, there is a food safety incidents database called "throwing out the window net" (http://www.zccw.info/) in China. It was developed by a group of volunteers in 2012 and it contains more than 3000 pieces of Chinese news about food contamination (Liu et al., 2015). The website editors have identified the food name and incident cause