Food Safety Warning Research Based on Internet Public Opinion Monitoring and Tracing

Hui Li, Hang Xiao, Tianchen Qiu, Pei Zhou*
1. School of Agriculture & Biology, Shanghai Jiaotong University
2. Key Laboratory of Urban Agriculture (South) of Ministry of Agriculture
3. Bor S. Luh Food Safety Research Center, Shanghai Jiaotong University
Shanghai, 200240, China
zhoupei@sjtu.edu.cn

***Abstract*—Internet has become an important platform of information publishing, exchanging and accessing. Recently, agricultural products and food safety relating news become hot posts on internet, which profoundly impacts daily life and public affairs management. The real-time and propagation characteristics of internet information makes internet public opinion becomes a more and more important food safety warning resource. In consideration of domain ontology has good concept hierarchical structure, includes abundant semantics, and has an outstanding significance on information resources organization and knowledge expression, food safety core ontology was built up in this paper. The food safety core ontology mainly focus on food safety incidents, invasive organism contamination, agricultural food sources pollution, processed food production issue, etc., based on semantic relations of concepts in ontology we designed inference rules, and achieved food safety knowledge inference and retrieval. Under the guidance of core ontology library, we designed self-adapting food safety internet hot public opinion identification and acquisition method, through customized crawler program, web pages were collected from internet and denoising processed, information document was generated from ontology library, followed by classification and realization of existing public opinion, those that can't be classified were computed by event dimension of vectors similarity for cluster analysis, and then updated the ontology library. All these effectively accomplish detect and trace food safety public opinions. In this paper, system of total merit index of food safety warning is constructed and the technique of quantitatively calculating different levels is given. Consequently, it contributes to scientific grounds for an objective, comprehensive and deep analysis in online information of food safety sector.**

Keywords—food safety; Internet public opinion; information extraction; vector space model ; early warning; evaluation index system

I. INTRODUCTION

Accompanied by the development of modern information technology, Internet has become an important carrier for information publication, exchange and acquisition. The multi-dimensional platform that brings to the public discussion have greatly shaped the nature of public opinion. For the past few years, information related to agro-product and food safety has been the focus of online discussion, deeply affecting everyday life and management of public affairs. Firstly, online information being real-time and transmissible makes it an important source for food safety warning, establishing reliable monitoring and analysis mechanism is therefore a must. Secondly, because of the real-time characteristic, exchangeability and openness of Internet, it is also required to build a mechanism that involves industry experts and government agencies to step in when needed. As a result, losses due to incidents caused by public food safety are minimized; preventing the spread of negative or even false information from being out of control [1].

This paper analyzes the real-time and transmissible characteristics of online information, based on the fast acquisition, content analysis, information tracing, model building and judgement of public sentiment of online information, establishing a food safety warning system upon Internet public opinion monitoring and tracing. Firstly, by building the core framework of the food safety sector, concepts in ontology are clarified in their literal relationships and logics. Secondly, by designing a self-adjusting technique for discriminating and collecting online popular discussion in food safety area, with the classification of information documentation vectors and realization of clustering algorithm, the monitoring and tracing of public opinions in known and underlying stages are effectively discovered. Lastly, total merit index system of food safety warning is built to provide scientific grounds to more objectively, totally and thoroughly analyze online public opinions in food safety sector.

II. STRUCTION OF FOOD SAFETY DOMAIN ONTOLOGY

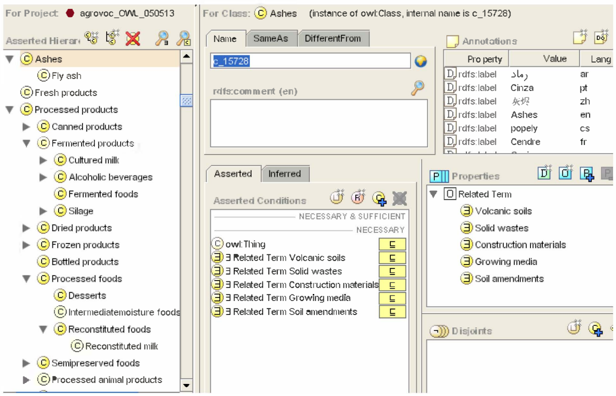
The so-called Ontology is the explicit and formal instruction of a shared conceptual model that can be explicitly and formally expressed to represent domain knowledge, improve mutual operability in a heterogeneous system and promote knowledge sharing. Specifically, Domain Ontology is an Ontology that has a specific description of object in a certain domain. The food safety domain sector that this thesis constructs mainly involves knowledge in food safety incidents, disqualified food products and toxic materials; includes explicit concepts and expressions among conceptual relationships.

Supported by Special Fund for Agro-scientific Research in the Public Interest of China (200903056), special funds for Shanghai Zhangjiang national independent innovation demonstration area(201209-SJ-A0-001) and the National High Technology Research and Development Program ("863" Program) of China (2012AA101405).

Upon a series of logical rules among designs of literal relationships that exist in domain concepts, it contributes to the semantic support for information abstraction and search in the food safety domain, and further achieves semantic search extension based on Ontology.

The structure of food safety domain Ontology is a 5-tuple [2] $O:=\{C,R,H^c,\mathrm{Re}l,A^o\}$。 The C and R here are two non-intersecting sets. The element C is called Concept; element R is called Relation ; H^c represents conceptual level which is the Taxonomy relation ; $\mathrm{Re}l$ represents Non-taxonomy relation in concepts; A^o represents Axiom. In the process of construction, technical details in acquisition of domain concepts, conceptual relations (includes Taxonomy relation and Non-taxonomy relation) and Axiom need to be determined.

The food safety domain Ontology described in this thesis involves unqualified food products, toxic materials and knowledge about food poisoning, which include category, attribute, relation, instance and other elements. Category is a named entity about problematic and toxic food products, also food safety incidents. Attribute includes time of occurrence, location, organization of production, toxic materials, classification and damage for food safety incidents. Relations include taxonomic relation and non-taxonomic relation. In this article, hyponymy relation is treated as taxonomic relation. Non-taxonomic relation refers to any logical relation other than is-a relation, such as has-a, synonymy, has-member-of, instance-of, is-member-of, is-part-of and has-part-of [3]. Cases are mainly in specific food safety incidents, problematic food products and toxic materials. Figure 1 shows the segmental instance of food safety domain ontology expressed by using OWL language.



For Project: agrovoc_OWL_050513

For Class: Ashes (instance of owl:Class, internal name is c_15728)

Name: c_15728

rdfs:comment (en)

Pro perty	Value	Lang
rdfs:label	رماد	ar
rdfs:label	Cinza	pt
rdfs:label	灰烬	zh
rdfs:label	Ashes	en
rdfs:label	popely	cs
rdfs:label	Cendre	fr

Asserted Conditions

NECESSARY & SUFFICIENT

NECESSARY

owl:Thing ⊑
∃ Related Term Volcanic soils ⊑
∃ Related Term Solid wastes ⊑
∃ Related Term Construction materials ⊑
∃ Related Term Growing media ⊑
∃ Related Term Soil amendments ⊑

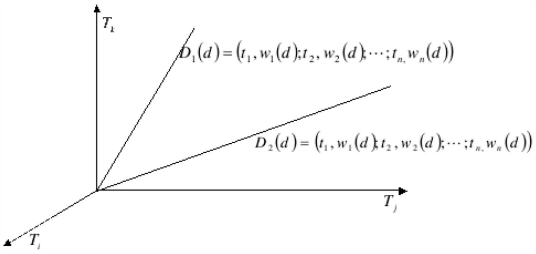
Fig. 1. Demonstration of food safety domain ontology fragments

III. METHODS OF INTERNET PUBLIC OPINION MONITORING AND TRACING

For Internet public opinion monitoring and tracing, it mainly involves the process of HTML files. VSM (Vector Space Model) is currently the most effective and practically used method, which is exactly the same model that is used to represent documentation in the research of this paper [4]. In VSM, documentation space is seen as a vector space consists of a set of orthogonal term vectors; that is, hypothetically if sum of features of all the documents is n , then an n -dimensional vector space is constructed. Among this space, every document is expressed as an n - dimensional characteristic vector. Consequently, after characteristics are extracted and weighted, document d could form an expression of characteristic space vector based on key words:

$$V_t(d):\{(t_1,W_1(d),(t_2,W_2(d),(t_3,W_3(d),\cdots(t_n,W_n(d))\}\ (1)$$

Among which, t_i is an entry item (vector), $w_i(d)$ is the weighted value of t_i in document d , $k=(1,2,\cdots n)$ is the number of entries to abstract for expressing the content of Web documents. We could consider $t_1,t_2,\cdots t_n$ as an n-dimensional coordinate system, and $W_1(d),W_2(d),\cdots,W_n(d)$ are the corresponding coordinate values. Therefore $V_t(d)$ is considered to be a vector in the n-dimension, called the vector expression of document d, as shown in Figure 2.



T_k
$D_1(d)=(t_1,w_1(d);t_2,w_2(d);\cdots;t_n,w_n(d))$
$D_2(d)=(t_1,w_1(d);t_2,w_2(d);\cdots;t_n,w_n(d))$
T_j
T_i

Fig.2 The document's vector space expressions

As textual data and format label are mixed up in HTML file, it makes the process of HTML file more complicated, and makes te most urgent problem to be solved is HTML file pre-processing, that is trying to make full use of useful information and abandon null noise [5]. This study used differential processing method that is similar to hyperlink.

TABLE I. WEIGHT OF DIFFERENT HTML TAGS

k	HTML tags	C_k
1	{Title}	8
2	{H1-H6、DL、OL、UL、U、I、B、Strong、EM、Table}	3
3	{A、Meta}	3/0
4	{else}	1

Besides HTML tags that represent HTML logistical structure information, HTML includes another information, that is webpage text language information [6]. The length of words is correlated with its expression ability, frequently used auxiliary, pronoun which have little help to reflect the document feature are less in length than those of noun, technical terms. Basically, the longer the word is , the less frequency it appears in the text, and usually has good expression and distinguishing ability, those traits make it to count the length of word string into consideration when calculating weight. Weight formula use word length weight modifying factor to represent word length effect to weight. The function value in different word length is shown as Table 2. As seen from the Table 2, the modifying factor has more weight reduction for short words, especial for single words. The auxiliary and pronoun which have little help for the document are mostly less than 2 in length. The weight of longer words such as proper noun and technical terms, was not greatly affected by modifying factor, which actually makes it relatively promoted, that is accord with the description of the concept form in domain ontology.

TABLE II. VALUES OF WEIGHT MODIFYING FACTOR BY WORD LENGTH

word length	0	1	2	3	4	5	6
modifying factor	0	0.63	0.87	0.95	0.981	0.993	0.998

The weighting scheme used in the system for webpage description was calculated using the following formula:

$$w_i = \frac{(1-e^L)\times\sum_k (C_k \times tf^k{}_i)\times\log(\frac{N}{df_i})}{\sqrt{\sum_{j=1}^{n}((1-e^L)\times\sum_k (C_k \times tf^k{}_i)\times\log(\frac{N}{df_i}))^2}} \qquad (2)$$

where $k \in \{1,2,3,4\}$ is the presence frequency of class k tag on webpage, L is the length of words string。 C is the corresponding weight of different HTML labels, w_i is the weight of i component on webpage vector. The formula was normalized to eliminate the effect of the document length. The method is based on ontology library in field to classify and cluster the public opinion information document vectors, it can not only effectively detect known public opinion information, but also identify the unknown information in initial stage.

IV. FOOD SAFETY WARNING EVALUATION INDEX SYSTEM

On account of public opinion collected, a objective evaluation index system is demanded to estimate the importance of the information and to trigger risk warning [7]. This paper constructed food safety internet public opinion monitoring index system, fully considered public opinion topical, information resources, communication channels, public opinion contents, audiences and macroscopic background, all of these factors come up to constitute a evaluation index set, and to certain extent realized non-materialized society phenomenon of public opinion in quantitative and qualitative analysis.

In the design of evaluation index system, following principles are obeyed: (1) Measurability. Due to lots of uncertain factors including in internet public opinion, the intended indicators should be workable, and can be represented quantitatively and qualitatively. (2) Reliability. The intended indicators should be able to make dependable and sensitive reaction when facing omen, be relatively stable and has self-updated function to guarantee the index system has continuity in certain time. (3) Traceability. Internet public opinion warning has the function to offer pertinent functional department with decision make services, thus, the intended indicators should be able to reflect current internet public opinion evolvement and developing trend. (4) Atomicity. The correlation of network public opinion index is complicated, and are reference for each other, consequently, a minimal complete index set is needed to be construct [8].

Based on the principles expatiate above, this study believes that internet public opinion starts from public opinion release source, and further on influent the audience. The effect that public opinion influent the audience is related to public opinion key elements and its macroscopic background. The paper listed 5 first level evaluation indexes, they are public opinion source index, public opinion key element index, public opinion audience index, public opinion dissemination index and public opinion macroscopic background index. Each first level evaluation index was elaborated to third level indexes that can be quantitative calculated. The basic weight of ach level of the index system can be calculated by analytic hierarchy process, and with the support of semantic analysis of ontology library in food safety area, every index weight was improved and optimized. The evaluation index system offers more objective, comprehensive and intensive scientific support for food safety internet public opinion analysis.

V. CONCLUSION

Based on the above processing method, by tracing and semantic analysis network public opinion on internet food safety news in 2012, a series of data was got that can essentially reflect the information point that need to be focused on. As in table 3, 14 key words that is most hot in food safety news after tracking 1089 hot topics in the relating network public opinion. Combined with manual analysis data, this food safety warning system is preliminary validated to be logical and feasible.

TABLE III. 14 MOST HOT KEY WORDS IN CHINA FOOD SAFETY INTERNET PUBLIC OPINION IN 2012

1	**2**	**3**	**4**	**5**
Illegal Cooking Oil	Food Safety	Edible Oil	Coca-Cola	Quality Supervision Bureau
6	**7**	**8**	**9**	**10**
Clenbuterol	Aflatoxin	Joincare	Tap-water	Additive
11	**12**	**13**	**14**	
Dairy Product	Bright Dairy & Food Cop.	KFC	Melamine	

Currently, it is quite requisite to study food safety warning technique that is based on internet public opinion monitoring and tracking. This paper uses the food safety domain ontology hierarchy, employs the internet hot public opinion identification and tracking algorithm to collect information, and has made certain progress, however, it is still on the development stage, and is a huge development space for its theory and application. In the process of follow-up system, precision and recall in information collecting need to be improved, evaluation index system needs to be consummate, experts participate level needs to be lower down, all of these are contribute to improve internet public opinion monitoring and tracking automation.

REFERENCES

[1] R.KDsala，H.Blockeel. Web Mining Research: A Survey. SIGKDD Explorations. 2000, 2(1):1-15.

[2] A. Maedche, S. Staab, Ontology Learning for the Semantic Web, IEEE Intelligent Systems, 16(2), March/April. 2001: 72-79.

[3] D.L. McGuinness, R. Fikes, Hendler J., et al, DAML+OIL: An Ontology Language for the Semantic Web, IEEE Intelligent Systems, 2002, 17(5): 72-80.

[4] J. Mahmud, H. Guo, A. Stent, I.V. Ramakrishnan. A general approach for partitioning web page content based on geometric and style information. In Submission to The 9th International Conference on Document Analysis and Recognition (ICDAR), 2007.

[5] S. Yu, D. Cai, J.R. Wen, et al. Improving pseudo-relevance feedback in web information retrieval using web page segmentation[C]//Proceedings of the 12th international conference on World Wide Web. ACM, 2003: 11-18.

[6] C. Bouras, V. Kapoulas, I. Misedakis. A Web page Fragmentation Technique for Personalized Browsing, ACM SAC 2004, Marc, 2004. M 14-17.

[7] Food and Drug Administration (FDA). Strategic Plan for Risk Commu nication. [2012-10-06]. http://www.fda.gov/AboutFDA/ReportsManuals Forms/Reports/ucm183673.htm.

[8] European Food Safety Authority (EFSA). Organisational structure. [201 2-10-10]. http://www.efsa.europa.eu/en/efsawho/efsastructure.htm.