2012 4th International Conference on Computer Modeling and Simulation (ICCMS 2012)
IPCSIT vol.22 (2012) © (2012) IACSIT Press, Singapore

Modelling Provenance in Food Supply Chain to Track and Trace Foodborne Disease

Qiannan Zhang[1],Dongyang Wang[1],Tian Huang[1],Yongxin Zhu[1],Meikang Qiu[2],Ming Ni[3],Guangwei Xie[3]

[1]School of Microelectronics, Shanghai Jiaotong University
[2]University of Kentucky
[3]No. 32 Institute of China Electronic Technology Group

Abstract. Food safety is important in large densely populated society. In this paper, a model of provenance in food supply chain is presented to provide strategies for stopping outbreaks of foodborne disease. We define a flexible data structure of provenance to organize related information of food supply chain so that algorithms can easily access them. Tracing and back-tracing algorithms are respectively implemented based on provenance to find out the pollutant source and problematical products downstream in the chain. We also propose a sampling strategy based on dynamic partitioning method to reduce the cost of sampling and improve the accuracy of algorithms. Simulation shows this model of provenance is effective to stop outbreaks of foodborne disease at low cost.

Keywords: Contamination, Foodborne disease, Provenance, Food supply chain, Strategy, Model, Simulation, Sampling, Tracing

1. Introduction

Rapid growth and industrialization have made the model of food supply chain move beyond regional and include global participation in importing and exporting in all levels of the chain According to the U.S. Census, the imported proportion of U.S. food consumption has grown from 7.9 percent to 9.6 percent between 1997 and 2005, roughly a 22 percent gain.[1] The limited capacity of existing regulations fails to monitor the food industry and no one group or a single response can remove all the food safety risks. The complicated structure of food supply chain provides pesticides, bacteria and drugs access to contaminate the food in each stage. To the extent that food ingredients are combined, processed and aggregated through a multitier supply chain, tracing processed foods in traditional way all the way back to the source of raw ingredient is extremely difficult. Now, information science is introduced to provide effective methods detecting the problematic spot which is responsible for the contamination in whole food supply chain.

Provenance can be represented in the form of meta-data that describes the ancestry or history of an object. It provides good properties for people to trace the origin of contamination when a food supply chain runs into problem. In this paper, we establish a model of provenance in food supply chain to solve the problem discussed above. Here are our contributions.

We first define a flexible data structure of provenance to organize related information of food supply chain.

In order to get a small and representative portion of samples from all products, we propose a sampling algorithm as the basis of tracing algorithm. This algorithm divides the data of products in unit of batch to make samples more general. In each partition, the sample rate is dynamic based on Bayesian Estimation to reduce the cost.

We present a tracing algorithm to detect the origin of contamination by traversing the entire history of the food supply chain. Back tracking algorithm is also implemented to find the potential contaminated food in the markets.

We simulate the model of provenance in several scenarios and find from the result that the model of provenance together with the algorithms we presented can effectively meet the requirements. It is indicated in simulation results that our DPS sampling scheme and algorithms can achieve up to tracing accuracy of 97.6% with an average sampling rate of only 8%.

The rest of this paper is organized as follow. Chapter 2 presents related work. Chapter 3 proposes the model of provenance and corresponding algorithms. Chapter 4 presents the simulation of the model and makes analyses on the result. Chapter 5 draws the conclusion.

2. Related work

In the context of food safety management, information systems are vital to assist decision making in a short time frame, potentially allowing decisions to be made in real time. McMeekin et al [5] introduced the technique of information systems used in the safety management of food supply chain. A stochastic state-transition simulation model [3] was described to simulate the spread of Salmonella from multiplying through slaughter, with special emphasis for critical control points to prevent or reduce Salmonella contamination. Wein et al [4] developed a mathematical model of a cows-to-consumers supply chain associated with a single milk-processing facility that is the victim of a deliberate release of botulinum toxin. Qin[2] established a quality management model for food supply chain based on game

3. Modeling

A provenance in food supply chain is viewed as a directed acyclic graph (DAG), in which each node stands for one location keeps some batches of foods for a period.

3.1. Data Structure

Traceability is the ability to trace the entire path of ingredients and food products from farms to factories to supermarkets' shelves. Rational data structure organization is the foundation of our design.

Considering the need of latter work, we defined two kinds of data structure in this chapter. All the two kinds of data will be stored in a central database.

3.1.1. Information of each location.

Table 1 shows the information recorded for every location. It includes the IDs of a batch passed through this location, the number of sampled products labeled as good or bad in this batch, and the IDs of polluted samples in this batch. Because the number of batches in one location is dynamic, linked list is chosen to store these data for its space-saving property and flexibility.

Table 1 data structure of location information

location	batch1	batch2	batch3	batch4	batch5	…
	bad1	bad2	bad3	bad4	bad5	…
	Good1	Good2	Good3	Good4	Good5	
	ID1_group	ID2_group	ID3_group	ID4_group	ID5_group	…

3.1.2. Information of every product

Table 2 records the location the food passed through during its procession.

Table 2 data structure of product information

ID	location1	location2	location3	location4	location5	……
bad_flag	batch1	batch2	batch3	batch4	batch5	……
	*p1	*p2	*p3	*p4	*p5	……

The order of "location" and "batch" forms the path of the product. "*p" points to address that store the information of corresponding location, which establishes the data communication between the two kinds of data structures. "bad_flag" displays whether the food product is healthy or not after testing.

3.2. Algorithms

The algorithms we proposed for model of provenance is consisted of three important parts: sampling from the whole collection of products; tracing the chain to detect the pollutant source; back tracking the network to get the potential contaminated products.

3.2.1. Sampling Algorithm

Since it is usually too expensive to test every product in the food supply chain, we propose Dynamic Partition Sampling Strategy (DPS) with the purpose to diminish the necessary number of sampling conducted before tracing and to retain the accuracy of our tracing algorithm.

The pseudo-code for Sampling Algorithm is shown as Fig. 1. Partition strategy can get more general and representative samples. In this case, the system partitions the whole group of products into several parts according to the batches they belong to in the end market. The sample volume for batch n of market m is determined by the formula:

$$sample\ number_{(m,n)} = sample\ number_{total} \times \frac{volume_{(m,n)}}{\sum_{M=0}^{m}\sum_{N=0}^{n} volumn_{(M,N)}} \quad \text{(Eq.2)}$$

Here, sample $number_{total}$ stands for the total number of samples in the network. Sample $number_{(m,n)}$ represents the samples in batch n of location m.

Then, dynamic strategy based on Bayesian Estimation is adopted to achieve minimal sample volume.[6] According to the infection probability of a particular pathogen got by medically experiments, the model can be trained to gain the distribution of total infection probability within the whole food supply network, which is called prior probability. The probability density function (pdf) of the distribution is presented as a function of infection probability intervals.

```
        Input: p: contamination probability according to the type
of contagion
        Get pollution probability distribution of the network
according to p;
        Sample n (small) products randomly;
        Compute posterior probability based on Bayesian
Estimation;
        Pick the interval with highest posterior probability;
        If  (pollution proportion < 80%)  {
```

Fig. 1 Pseudo-code for Sampling Algorithm

On the other hand, after sampling a small part, the infectious rate of the samples, as the posterior probabilities can be obtained.[7] If k tainted products are found within n samples, under each contamination percentage interval with prior probability of a_i%, conditional probability is got by binomial distribution:

$$P(B \mid A_i) = \binom{n}{k} a_i^{\,k}(1-a_i)^{n-k} = \frac{n!}{k!(n-k)!} a_i^{\,k}(1-a_i)^{n-k} \quad \text{(Eq.3)}$$

Here, A_i means the case that the contamination percentage of the whole products falls into the interval with prior probability of a_i% and B means the case that we find k contaminated products in n samples.

After that, Bayesian Formula is applied to mix prior probabilities with posterior probabilities to get more precise probabilities.

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \quad \text{(Eq.4)}$$

Consider about the characteristics of the tracing algorithm, which will be discussed in the next section, there are some requirements for certain circumstances. If the ratio of bad products is too high, the tracing algorithm performs poorly since there are not enough healthy samples to exclude the suspicions. On contrary,

if the ratio is too low, the noise introduced by sampling process may dominate the result. In these two cases, more samples should be tested to improve the accuracy. Hence, under each sample rate, there is a relationship between tracing algorithm accuracy and pollution proportion interval and normally.

With all the relationships in a specific interval, economically, pick up the smallest sample rate that achieves certain requirements (like 90% accuracy). Sample again under that rate and update Bayesian Estimation to find if the tracing accuracy meets the requirements.

3.2.2. Tracing Algorithm

After sampling and testing, the system checks the information stored for every sampled product and finds the corresponding locations and batches it passed through to get the "good" and "bad" variables which stand for the number of sampled healthy and contaminated products. The pseudo-code of tracing algorithm is shown as Fig. 2.

```
for (all m∈Samples) {
if (m is polluted) bad += 1 for all locations and batches;
        else       good += 1 for all locations and batches
of the product;}
find all batches with good < ε && bad >0;
if (only one suspect)   output the suspect spot as origin;
else {
get food IDs passed all suspects;
        construct "Suspect Tree" of batches according to
the IDs' paths;
        breadth-first traverse the tree to find the first node
```

Fig. 2 Pseudo-code for tracing algorithm

Suppose that the samples can properly map the condition of the whole group of products, thus, "good" and "bad" variables can reflect the proportion of unpolluted and polluted products in entire end market. The criterion of the judgment to find the suspects is "good" < ε and "bad" > 0. That's because the pollutant source will be the primary spot generates polluted food while the number of healthy samples is limited there. ε, considered as the error factor, is a small integer which can make the algorithms still valid when part of the food passed the source is unpolluted or there is disturbance of the non-ideal problems(e.g. imperfect sampling). The specific ε value is decided by the samples' number and infection probability of pollutant source, which can be roughly represented as:

$$\varepsilon = \frac{\text{sample's number} \times \text{pollution probability}}{\text{number of batches}} \quad \text{(Eq.1)}$$

In this way, more than one suspect may be found. More work should be applied to eliminate the confusion suspects. First, in order to improve the speed of the algorithm, suspects with small "bad" value will be excluded. Then, the system will generate a "Suspect Tree" composed of the suspected locations and batches according to their order in food supply chain. After that, traverse the "Suspect Tree" layer by layer and since the real contaminant is on the upper hierarchy over others, the first node that meets the criterion will be picked up.

3.2.3. Back Tracking Algorithm

The last task is to back track all infected products in the end market so as to prevent these harmful foods from further endangering people's health.

In order to judge the performance of back tracking algorithm, Hit Rate and False Alarm Rate are introduced to denote the ability to capture polluted products and the probability of selecting good products by mistake. Supposing the total number of bad products is B and the algorithm selected n products, including b polluted ones and g unpolluted ones, we define Hit Rate as b/B, and False Alarm Rate as g/n. Although, theoretically, both high Hit Rate and low False Alarm Rate are expected, there is a tradeoff between them. Nevertheless food safety issues are critical problems, Hit Rate should be given priority and False Alarm Rate could be sacrificed when necessary.

The back tracking algorithm is described as follows:

Exam the samples and find the infected products.
Construct a tree of locations and batches according to products' paths.
Traverse the tree (DFS or BFS) from pollutant source and record all the locations and batches.
Select all the products that passed those recorded spots.
If there are new locations and batches after the traverse, update the bad products and go to 0.

4. Simulation

4.1.Data set

In order to simulate the model of provenance we proposed, an example of data set for a food supply chain network is presented as Fig. 3.

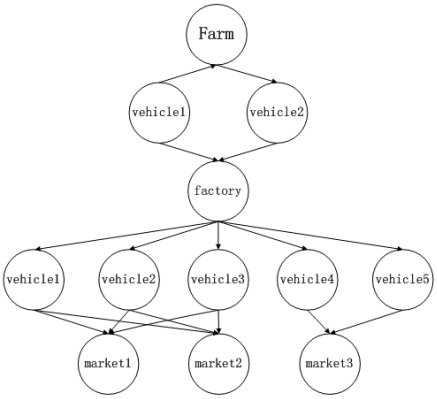

Fig. 3 The Topology Structure of the chain

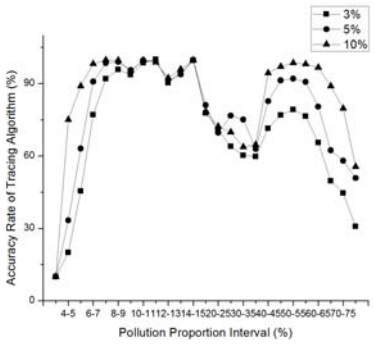

Fig. 4 The relationship of tracing algorithm accuracy and pollution proportion intervals under different sample rates

In this case, location reusing is considered. Suppose every location has 10 batches. 6k products with unique IDs randomly pass through these locations.

To simulate the impact of origin, the pollutant source is first set by the system. Food will be infected (1) by pollutant source directly or (2) by sources of cross infection indirectly according to Reed Frost Model.

4.2. Prior information

The system trains the specific model to get prior probabilities. Fig. 5 shows the distribution of pollution proportion under different contamination probability after 30k tests. The relation between the products' contamination percentage and the tracing algorithm accuracy under different sample rates is shown as Fig. 4.

For simplicity, only 3 sampling rates are tested: 3%, 5% and 10%. To make sample strategy more efficient, more sampling rates can be tested.

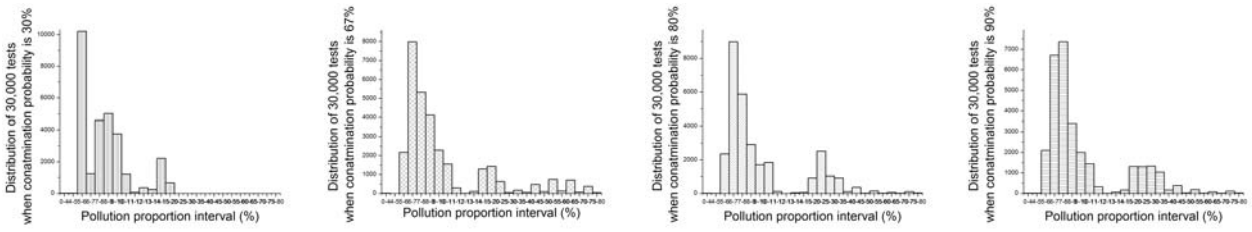

Fig. 5 Prior Probabilities distribution under different contamination probability

4.3. Results

For tracing algorithm part, the system is tested under different probabilities of contamination. The accuracy of the algorithm is listed in Table 3

Table 3 Simulation results of tracing algorithm

probability of infection	Accuracy
30%	80.8%
67%	92.6%
80%	97.6%
90%	95.8%

For sampling algorithm part, we first compared partition sampling with global sampling. Table 4 shows the accuracy of tracing algorithm.

Table 4 Simulation results of partition sampling algorithm

probability of infection	global random sampling	partition random sampling
30%	56%	67.8%
67%	61.5%	84.2%
80%	74.5%	85.4%
90%	82%	86.1%

Then, dynamic method is tested to further approve the advantage of our sampling algorithm. The results are presented in Table 5.

Table 5 Simulation results of dynamic sampling algorithm

probability of infection	3% sampling proportion	5% sampling proportion	10% sampling proportion	dynamic sampling
30%	67.8%	78.4%	88.5%	80.8%
67%	84.2%	92.3%	91.6%	92.6%
80%	85.4%	92.4%	96.6%	97.6%
90%	86.1%	92.4%	93.0%	95.8%
average sampling proportion	3%	5%	10%	8%

For back tracking part, the result of simulation is shown as Table 6.

Table 6 Simulation results of back tracking algorithm

Hit Rate	False Alarm Rate
96%	56%

The results of the simulation show that the accuracy of tracing algorithm is satisfying. However, in some cases, the algorithm cannot correctly detect the pollutant source. The reasons are as follows:

The samples are not strictly uniformly distributed.

During traverse of the “Suspect Tree”, in the same stage, the system may get to a false spot satisfies the criterion first then get to the origin.

5. Conclusion

In this paper, we propose a model of provenance in food supply chain to provide strategies for stopping outbreaks of foodborne disease. We present a tracing algorithm to find the source of contamination of a food supply chain. We propose Dynamic Partition Sampling Strategy to reduce the cost of sampling. Besides, we also propose a back-tracing algorithm to provide strategy for recalling problematical food undiscovered in the chain. It is indicated in simulation results that our DPS scheme and algorithms can achieve up to tracing accuracy of 97.6% with an average sampling rate of only 8%.

In this paper, We assume that all provenance information of food products is held in a centralized database and these provenance meta-data are orgnized in a uniform manner. Our future work is to further make practical implementation of the provenance of food supply chain based on cloud storage services.

6. Reference

[1] Roth, Aleda V.,Tsay, Andy A.,Pullman, Madeleine E.,Gray, John V. Unraveling the Food Supply Chain: Strategic Insighents from China[J]. Journal of Supply Chain Management. 2008.

[2] Qin Li; Wang Qing Song. Food Supply Chain Quality Management Model and Simulation Based on Game. International Conference on Computer Modeling and Simulation, 2009. ICCMS '09. PP:291 – 293

[3] Monique A. van der Gaag, Fred Vos, Helmut W. Saatkamp, Michiel van Boven, Paul van Beek, Ruud B.M. Huirne, A state-transition simulation model for the spread of Salmonella in the pork supply chain, European Journal of Operational Research, Volume 156, Issue 3, 1 August 2004, PP: 782-798

[4] Wein, LM; Liu, YF. Analyzing a bioterror attack on the food supply: The case of botulinum toxin in milk. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA. 2006 Volume: 102 Issue: 28 PP: 9984-9989

[5] McMeekin, TA; Baranyi, J; Bowman, J; Dalgaard, P; Kirk, M; Ross, T; Schmid, S; Zwietering, MH. INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY. 2006 Volume: 112 Issue: 3 PP: 181-194

[6] Zhang Heguan. Empirical Bayes methods applications in seedlings [J]. Beijing: Eighty-one Agriculture College. 1986.

[7] Chen ping, Hou Chuanzhi, Feng yuyu. Random Math [M]. Beijing: National Defence Industry Press. 2008: 165-187.