A study on association rules-based mining of consumer product accident information

Yingcheng Xu
Quality Research Branch
National Institute of Standardization
Beijing 100191, China
15810265898@126com

Xiuli Ning
Quality Research Branch National
Institute of Standardization
Beijing 100191, China
nxl_warm0908@163.com

Ya Li
Quality Research Branch
National Institute of Standardization
Beijing 100191, China

***Abstract*—Applying association rules to mining of consumer product accident information, this paper demonstrates the processes for generating item set, transaction set and candidate set, and studies 415 cases of toy products. The results indicate the approach is both feasible and reasonable. The approach may provide technical guidance for mining and analysis of consumer product accident information.**

Keywords—association rules, consumer product, information

I. INTRODUCTION

The quality and safety of consumer products remains closely related to the health of general public, healthy development of economy as well as social harmony and stability. Affecting national economy and people's livelihood, it is not only a common concern of consumers but also a focus for government, business, academic circle and public opinion. The information about consumer products' quality and safety features high diversity, wide coverage, large size, and quick change. Moreover, the accident information concerning the quality and safety of consumer products is quite concealed in the early stage. But once reaching a certain point of time, it can be quickly and widely spread with a turning point that is really hard to be captured. The resulting issues can hardly be resolved through traditional statistical analysis. There is a paucity of studies focusing on the mining and analysis of consumer products' quality and safety accidents in our country. Therefore, it is an inevitable option to develop studies about consumer product accident information mining and improve the abilities of supervising, determining, alerting and disposing the information concerning quality and safety risks of consumer products.

II. ASSOCIATION RULES-BASED CONSUMER PRODUCT ACCIDENT INFORMATION MINING PROCESS

The association rules-based consumer product accident information mining process mainly covers the ways of generating three sets, namely item set, transaction set and candidate set.

A. Item set generation

In this paper, each case is viewed as a transaction and the combination of field and its value (namely "field-field value" pair) as an Item. Suppose there are *m* fields in the case in total, p_i means no. *j* field of the case, and $|p_i|$ means the number of p_i 's values, $|p_i|$ items can be generated from field p_i and they are $I(p_i) = \{I_{i1}, I_{i2}, \ldots, I_{i|p_i|}\}$ where $I(p_i)$ indicates the collection of items generated on the basis of field p_i and $I_{ij} = < p_i, v_{ij} >$ (j =1,2,…, $|p_i|$)is no. *j* item in $I(p_i)$, and v_{ij} is no. j value of field p_i. In this way, a set of all the items can be obtained, $I = \{I_1, I_2, \ldots, I_m\} = I(p_1) \bigcup I(p_2) \ldots \bigcup I(p_m)$. Corresponding algorithm description is as follows:

Algorithm 2.1 Item set-generating algorithm

1: Initialize, set *i*=1;
2: Read all the possible values of $|p_i|$ fields for no.*i*field $v_{i1}, v_{i2}, \ldots, v_{i|p_i|}$;
3: Add $I_{ij} = < p_i, v_{ij} >$ (j=1, 2, ..., $|p_i|$) into $I(p_i)$ in turn and generate $I(p_i) = \{I_{i1}, I_{i2}, \ldots, I_{i|p_i|}\}$;
4: Set $I = I \bigcup I(p_i)$ and $i = i + 1$;
5: If $i < m$, go to line 2;
6: Output the item set *I*.

B. Transaction set generation

For the numeric fields, their values may undergo interval division first and then index number of original field value's interval should be used as the new field value. And for text fields, text categorization algorithm can be applied to the original text to complete categorization and the resulting text categories can be used as the new values for the fields. When the text in certain text field of a case is identified as belonging to several categories after the categorization, the field of this case will be deemed as having multiple values which should be separated with "##" when stored, e.g. "Category 1##Category 3##Category 5". Suppose there are N cases in total and they are saved as $CASE = \{case_1, case_2, \ldots, case_N\}$, the algorithm for generating the transaction set $T = \{ T_1, T_2, \ldots, T_N \}$ is depicted below:

Algorithm 2.2 Transaction set generating algorithm

1: Initialize, set k =1;

2: Read the information about no. k case $case_k$;

3: Set $i = 1$;

4: Read the value of no. i field pi in $case_k$;

5: Deal with the case as per the field value of p_i :

7: ①When value of field p_i in $case_k$ is empty, go to line 10;

8: ②When the number of values of field p_i in $case_k$ is 1, read field value v_{ij} and add $I_{ij} = < p_i, v_{ij} >$ into T_k ;

9: ③When the number of values of field p_i in $case_k$ is higher than 1, read each field value v_{ij} in turn and add $I_{ij} =< p_i, v_{ij} >$ into T_k ;

10: $i = i + 1$;

11: When i is lower than the number of all the fields, go to line 4;

12: $k = k + 1$;

13: When $k \le N$, go to line 2;

14: Output transaction set T ={ $T_1, T_2, ..., T_N$ }.

C. Candidate set generation

Based on level-wise search thought, "downward closure" principle is used for combining and pruning purposes so as to compress the search space when generating candidate set. In present study, the item is a combination of "field and field value". Only some text fields may have several field values, all other fields have only one field value (being empty sometimes) in one case. In view of this, in generating candidate set, the items from same field can't be combined (except for text fields). For instance, in one candidate set, items "sex-male" and "sex-female" can't be contained at the same time. However, some text fields may have more than one values (or belong to different categories at the same time). In order to distinguish such situations, $Multi(p_i)$ is used to identify whether field p_i has multiple values. If field p_i can have more than one values, $Multi(p_i) = 1$; or else, $Multi(p_i) = 0$. Corresponding algorithm is as follows:

Algorithm 2.3 Improved candidate set generating algorithm based on Apriori

Input: ($k-1$)-frequent item set F_{k-1}

Output: k -candidate set C_k

1: Set $C_k = \phi$;

2: Start the circulation;

3: Search the frequent set pair (f_1, f_2) that has "unique last item" in F_{k-1} . The set pair meets the following conditions: $f_1 = \{I_1, ..., I_{k-2}, I_{k-1}\}$, $f_2 = \{I_1, ..., I_{k-2}, I'_{k-1}\}$, and $I_{k-1} < I'_{k-1}$, $f_1, f_2 \in F_{k-1}$.

4: Figure out which fields and their attributes generate items I_{k-1} and I'_{k-1}, set the fields corresponding to I_{k-1} and I'_{k-1} to be p_i and p'_i, respectively.

5: If p_i and p'_i are different from each other, or they are the same as each other and $Multi(p_i) = 0$, combine f_1, f_2 , namely $c \leftarrow \{I_1, ..., I_{k-2}, I_{k-1}, I'_{k-1}\}$; or else, go to line 3 to find out next frequent set pair that "has unique last item" in F_{k-1} ;

6: $C_k \leftarrow C_k \cup c$ 。

7: Go through all the ($k-1$)-item set s in c, if $s \notin F_{k-1}$, eliminate s from c;

8: Keep circulating until finding all the frequent set pairs (f_1, f_2) that have "unique last item" in F_{k-1} ;

9: Output C_k .

III. APPLICATION AND ANALYSIS

A. Data format

In this paper, altogether 415 cases of toy products are selected. The data format basically meets the requirements posed by association rules-based mining. Some data are illustrated in Table I.

TABLE I CASES OF TOY PRODUCTS

No.	Case
1	Toy car; coating; lead; exceeded; intoxication
2	Rattle; toy; tiny component; suffocation
3	Mickey Mouse; toy; sounding device; suffocation
4	Duck; toy; whistle; suffocation
5	Swim ring; toy; DINP; intoxication
6	Toy rattle; ball; suffocation
7	All-terrain vehicle (ATV); toy; front wheel brake; tire gauge; loss of control; fall
8	Storage box; toy; cover; trapped
9	RC helicopter; toy; battery; burn
10	Magnetic dart; toy; tiny magnet; fall-off; fatal
11	Cinderella; electric car; toy; short circuit; fire accident
12	Lego; toy; lead; intoxication
13	Toyracing car; lead; intoxication
14	Magnetic assembly; toy; tiny magnet; fatal
15	Toy pacifier; volume; suffocation
16	Metroid; toy; lead; intoxication
17	Frog block; toy; lead; intoxication
18	Toyhorse; coating; lead; exceeded; intoxication
19	Small Rider toy; coating; lead; exceeded; intoxication
20	Rattle; toy; tiny component; suffocation
21	House; toy; tiny component; suffocation
22	Bear; toy; fall-off; suffocation
23	Whale; toy; DEHP; intoxication
24	Toy; oil paint; lead exceeded; intoxication
25	Cosmetics; toy; chromium; lead; intoxication
26	Car; toy; chromium; lead; intoxication
27	Alarm instrument; toy; tiny component; suffocation; edge; scratch
28	Airplane; toy; tiny component; suffocation; rope length
29	Toyblock; tiny component; suffocation; edge; scratch
30	Mobile phone; toy; sound; acoustic trauma
31	Expansion egg; toy; expansion rate; suffocation
32	Mobile phone; toy; sound; acoustic trauma
33	Crack Troops; toy; lead; intoxication
34	Gardening; toy; lead; intoxication
35	Crystal ball; toy; lead; intoxication
36	Fishing game; toy; lead; intoxication
37	Mobile phone; sound; acoustic trauma
38	Car; chromium; lead; intoxication
39	Plush doll; toy; lead; intoxication

B. Outcomes of algorithm analysis

Mining of toy product cases is performed using association rules at 0.07 support degree and 0.6 confidence level to obtain corresponding results (see Table II).

TABLE II ASSOCIATION RULES-BASED MINING OUTCOMES

Frequent item set	**Support degree**	**Rules**	**Confidence**
Lead	0.1932367	Tiny component-->suffocation	0.9873418
Intoxication	0.4202898	DEHP-->intoxication	1
Suffocation	0.3599034	DINP-->intoxication	1
Tiny component	0.1908213	Lead-->intoxication	0.975
DINP	0.07246377		
DEHP	0.120773		
Mobile phone	0.07246377		
Intoxication, lead	0.1884058		
DINP,intoxication	0.07246377		
DEHP,intoxication	0.120773		
Tiny component, suffocation	0.1884058		

As revealed in the table, theprobability for [lead, intoxication] is about 18.84%, that for [DEHP, intoxication] is 12.07%, that for [tiny component, suffocation] is 18.84%. Among all the cases of lead-contained consumer products, the probability for intoxication is 97.5%; among all the cases of consumer products containing tiny components, the probability for suffocation is 98.7%; and among all the cases of DEHP (Di 2-Ethyl Hexyl Phthalate)-contained consumer products, the probability of intoxication is almost 1.

For lead–>intoxication rule, lead is a heavy metal element that can cause severe impact on human health from the medical perspective. Children remain especially vulnerable to lead poisoning. When being exposed to lead-containing toys, lead will enter human body through respiratory and digestive tracts and get deposited in bones and then flow with blood to reach all the organs in the body. It can cause such toxic reactions as diarrhea, anemia and vomiting, and severely affect the development, mentality and growth of children. Therefore, toy manufacturers and inspection & quarantine institutions should strictly control the lead content in the toys. The rule of tiny component–>suffocation is also a problem that is usually ignored by different toy manufacturers. Since children are lively and active in nature, when tiny components in the toys fall off naturally or due to artificial force, they may be swallowed accidently by children. This can easily cause risk of suffocation.

IV. CONCLUSIONS

Based on a library of consumer product accidents, this paper employs algorithms of association rules to analyze potential relations among type and degree of harm as well as consumer product factors in related cases. It can provide alert to the safety of consumer products and thus has significance for practice.

ACKNOWLEDGMENT

We would like to acknowledge that this research is supported and funded by the National Science Foundation of China under Grant No. 91646122, 91746202 and 71571191, the Basic Scientific Research Business Projects 552018Y-5927-2018.

REFERENCES

[1] Han J., Kamber M. Data mining: Concepts and techniques. New York: Morgan Kaufmann Publishers Inc., 2006

[2] Yang Y., Liao Y., Meng G., et al. A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. Expert Systems with Applications, 2011, 38(9): 11311-11320

[3] Chang J. H. Mining weighted sequential patterns in a sequence database with a time-interval weight. Knowledge-Based Systems, 2011, 24(1) :1-9

[4] Huang T. C. Mining the change of customer behavior in fuzzy time-interval sequential patterns. Applied Soft Computing, 2012, 12(3) :1068-1086

[5] Kaneiwa K., Kudo Y. A sequential pattern mining algorithm using rough set theory. International Journal of Approximate Reasoning, 2011, 52(6) :881-893

[6] Schedl M., Widmer G., Knees P., et al. A music information system automatically generated via Web content mining techniques. Information Processing & Management, 2011, 47(3): 426-439

[7] Mykowiecka A., Marciniak M., Kupsc A. Rule-based information extraction from patients' clinical data. Journal of Biomedical Informatics, 2009, 42(5): 923-936