A case study on integration of vehicle defect and recall information analysis

Yasuyuki Yamamoto*1, Shu Yamada1

1Keio University, Hiyoshi 3-14-1, Yokohama, Kanagawa

*yamamotoyasuyuki100@keio.jp

Abstract:
As the information of vehicle defect, the vehicle defect information and the recall information are disclosed by the website of Japanese Ministry of Land, Infrastructure, Transport and Tourism. Unlike other studies which analyze the text data of each one, we focus on the relation between them and analyze it of both. In order to derive the tendency of dependency pairs which means defect phenomena, dependency parsing and BM 25 are applied for the analysis. From results of calculating occurrence and BM 25 score for pairs of dependency words and modified words, some tendencies are found in each of high score’s pairs. For the analysis of recall information, correspondence analysis, co-occurrence network and self-organizing map are conducted to find the relational words between three categories of defect factor. Twelve primary factors of five parts are shown as the factors in fuel device which causes liquid leakage. Finally, factors of important defect phenomena is discussed. Several dependency pairs are shown as important defect phenomena, and primary factors of parts are shown as an example of factors of liquid leakage which is one of them.

Keywords
Text mining; self-organizing map; correspondence analysis; co-occurrence network

1. Introduction

As the information of vehicle defect, the vehicle defect information and the recall information are disclosed by the website of Japanese Ministry of Land, Infrastructure, Transport and Tourism. Previous researches analyze each ones and derives results for prevention. In this research, from understanding type and content of each text data, different methods are applied for analysis of both. As a result, pairs of words which mean defect phenomenon from analysis of the vehicle defect information and relative combination of three words of defect factor from the recall information is derived. Finally, to derive the information which is combined defect phenomenon and defect factors, integrating analysis with results of them by proposal way is conducted. As a result, the information is derived as factors of the important defect phenomena.

2. Previous research and focus point of this research

2.1 Analysis of text data of vehicle defect
Nishiura and Yamada (2008) extracted the vehicle defect information which was related to top three Japanese companies in passenger car sales in 2004 (Toyota, Nissan, Honda), and analyzed the words of “summary of report” in it. First, typologizing the defect information and analyzing was conducted by morphological analysis and KeyGraph map which could show the relational pair of words made by Ohsawa et al. (1999), and the relation of defect words between “system” and “phenomenon” was derived. After that, analyzing the relevance between of defects was conducted by correspondence analysis and association rules. Finally, as the feedback for selecting items at design review, a new format was proposed by reflecting the results.
Kitazawa and Osada (2008) used the recall information of four major car companies from 1993 to 2006, and analyzed them with 4 viewpoints: tendency of recent recall, defect situation, secular change of defect factor, and impact of defect. In the chapter of “defect situation” which was focused on the text data of “situation” in it, by aggregating the numbers of three categories of defect factors (defect systems, parts and primary factors), considerations of each category for each companies were derived.

2.2 Focus point of this research

From the investigation into the vehicle defect information, it is found that almost includes expressions of defect phenomenon which is described as the phrase of “system” and “phenomenon” which can be found by not well versed users. From here, this research focus on defect phenomenon. These two words of defect phenomenon has a relationship of dependency. Therefore, this research analyzes these two words not look like one by one, but look like connected with relationship.

The recall information has contents which are written in fixed phrases because it is reported by company. Among them, this research focus on defect factors because it is considerable that there is no other information which describe factors accurately and it can be useful for kaizen works. In order to analyze it for kaizen works, it is better to show categories of defect factors with combinations. Therefore, this research analyzes it for deriving strong combinations of categories of defect factors.

It is better to show defect factors which causes important defect phenomenon. From this consideration, by integrating these two results, deriving factors of important defect phenomena is conducted.

3. Analysis of the vehicle defect information and quantitative evaluation for defect phenomena

The vehicle defect information can be searched from the website of Ministry of Land, Infrastructure, Transport and Tourism (http://carinf.mlit.go.jp/jidosha/carinf/opn/index.html). For analyzing, 5768 data are extracted with under following conditions.

- Period: from 1st April 2008 to 31st March 2013
- Company: Toyota, Honda, Nissan
- Content: One with only one comma in "Summary of report content"

So as to evaluate pairs of two words which has a relationship of dependency, dependency parsing and BM 25 are applied. Dependency parsing is one of the methods which analyze a grammatical construction of sentences. Dependency pair which is made of modifier words and modified words and each numbers are derived. It is considerable that the number of dependency pairs which means defect phenomenon can indicate the number of occurrences of defect phenomena which the pair means. BM 25 is a formula which calculate a strongness of the relationship between the sentences and each word by frequency, narrows of spreading and shortness of letters in sentences which include the word. In order to get the information of totally importance which is cannot found by only the numbers, analyzing by this for dependency pairs and modified words is conducted.

3.1 Deriving the occurrence of the defect phenomenon by dependency analysis

So as to conduct dependency parsing, under the following conditions for analysis are set.

- Software: TTM (TinyTextMiner)
- Forbidden Words: seven words (“Naru”, “Suru”, “Iru”, “Yoru”, “Niru”, “Reru”, “Aru”) which impede the distinctness of dependency pairs because of no meaning and extensive existence

As a result of dependency parsing, 15960 dependency pairs and each number is derived. Table.1 shows the comparison with top ten occurrence pairs and ten pairs which the number of occurrences is two which are extracted by random sampling. From investigation of meanings whether defect phenomenon or not, it can be found that most of top 10 occurrence pairs mean defect phenomenon, but the others are not so much. From here, it is considerable that there is a tendency which means defect phenomenon on high occurrence pairs.

3.2 Deriving importance of defect phenomenon and effect of phenomenon by BM 25

At the analysis by BM 25, BM 25 score is shown as a result. In this research, BM 25 score is defined as the numbers which is calculated from the formula which is reference from Yoshioka and Koeda (2012).

\[
score(w, D) = \sum_{i=1}^{n} idf(w) \cdot f(w, d_i) \cdot (k_1 + 1) / (f(w, d_i) + k_1 \cdot (1 - b + b \cdot |d_i|/avgd_i))
\]  

(1)
In addition, $N$ is the total number of sentences, $n(w)$ is the number of sentences containing $w$, and $f(w, d_i)$ is the number of occurrences of $w$ in $d_i$, divided by the total number of words in $d_i$. $k_1$ and $b$ are parameters of BM 25, and we adopt $k_1 = 2.0$ and $b = 0.75$ that is widely used in BM 25. Besides conditions of BM 25, condition of dependency pair is set as from the result of dependency parsing and condition of modified word is set as only verbs and adjectives because these are expected as expressions of phenomenon. With these conditions, analyzing dependency pairs and modified words from the result of dependency parsing is conducted. Table 2 shows the top 10 pairs of BM 25 score of dependency pairs.

Table 1. Comparison with top 10 occurrence pairs and 10 pairs that occurrence is 2

| Dependency pairs | Occurrence | Meaning | Dependency pairs | Occurrence | Meaning |
|------------------|------------|---------|------------------|------------|---------|
| Engine + stop    | 332        | Findable defect | Unstable + work | 2          | Incoherent |
| Transmission + defect | 167 | Defect factor | Gasoline + change | 2 | Incoherent |
| Accel pedal + step | 157 | Add “not” then defect phenomenon | Lever + down | 2 | Add “not” then defect phenomenon |
| Alert light + light | 111 | Add “not” then defect phenomenon | Water + pour | 2 | No meaning |
| Sound + occur    | 97         | Findable defect phenomenon | Moving + extend | 2 | Incoherent |
| Running + stop   | 94         | Findable defect phenomenon | Alert sound + occur | 2 | Add “not” then defect phenomenon |
| Engine + work    | 91         | Add “not” then defect phenomenon | Stop + move | 2 | Incoherent |
| Running + light  | 87         | Add “not” then defect phenomenon | Front glass + reflect | 2 | Normal phenomenon |
| Break + work     | 82         | Add “not” then defect phenomenon | Running + bite | 2 | Incoherent |
| Engine + start   | 56         | Add “not” then defect phenomenon | Oil + circulate | 2 | Add “not” then defect phenomenon |

Table 2. Top 10 pairs of BM 25 score of dependency pair

| Dependency pairs | BM 25 score of dependency pair | The number of changing from dependency parsing |
|------------------|--------------------------------|-----------------------------------------------|
| Engine + stop    | 74.0                           | 0                                             |
| Running + stop   | 27.4                           | 4                                             |
| Transmission + bad | 26.8                   | -1                                            |
| Engine + start   | 23.6                           | 3                                             |
| Abnormal Sound + outbreak | 20.0                    | 0                                             |
| Alert light + light | 19.2                    | -2                                            |
| Oil + leak       | 18.8                           | 5                                             |
| Fuel + leak      | 18.2                           | 5                                             |
| Accel pedal + step | 18.0                    | -6                                            |
| Brake + work     | 15.4                           | -1                                            |

From checking BM 25 score of dependency pair, it can be found that only “Engine + stop” scores bigger number but score of the others do not make such a difference from others. From considering the distribution of occurrence from Table 1, it can be said from the macro perspective that occurrence has bigger effect in the analysis of BM 25 for dependency pairs which has high occurrence than narrows of spreading and shortness of letters in sentences.
For checking results from micro perspective, the number of changing from dependency parsing is calculated. From here, it can be found that some changing of rank has occurred, and the pairs which mean critical defects like “oil + leak” and “fuel + leak” score positive number but the pairs which do not mean like that directly like “accel pedal + step” and “alert light + light” score negative number. Because of them, it can be said that this score shows importance of defect phenomenon which the dependency pair means.

Table 3 shows the top 10 pairs of BM 25 score of modified word and expectable dependency pairs selected by top occurrence. From here, it can be found that the words which mean serious phenomenon leading to the outage like “leak” and “occur” are placed top two and the downer the word place, the smaller effect which the word means. From this tendency, it is considerable that BM 25 score of modified word would show importance of effect.

### Table 3. Top 10 pairs of BM 25 score of modified word

| Modified words | BM 25 score of modified words | Expectable dependency pairs |
|----------------|-------------------------------|-----------------------------|
| Leak           | 39.2                          | Fuel + leak                 |
| Occur          | 37.1                          | Clack + occur               |
| Work           | 26.0                          | Break + work                |
| Step           | 21.9                          | Accel pedal + step          |
| Occur          | 18.0                          | Clack + occur               |
| Work           | 16.4                          | Engine + work               |
| Open           | 14.1                          | Door + open                 |
| Bad            | 13.8                          | acceleration + bad          |
| Decrease       | 12.1                          | engine oil + decrease       |
| Up             | 12.0                          | rev + up                    |

### 4. Revealing the related combinations of defect factors from the analysis of the recall information

The recall information also can be searched from the website of Ministry of Land, Infrastructure, Transport and Tourism [http://carinf.mlit.go.jp/jidosha/carinf/ris/index.html](http://carinf.mlit.go.jp/jidosha/carinf/ris/index.html). In this research, 1038 data are extracted with the under following conditions.

- **Period:** from 15th April 1993 to 17th December 2018
- **Targeted manufacturer:** Toyota, Honda, Nissan
- **Extracted content:** “Defective system” and “Situation” that excluded unknown factor and complex factors

Data cleaning is conducted, because it is written in fixed phrase, the part of defect factors is positioned in same place in almost all of sentences. Extracting from top to the word of “Tame” where it is positioned is conducted for sentences which is written in fixed phrase. For other sentences which do not exist “Tame”, extracting before “Mono ga aru” or “Koto ga aru” or “Osore ga aru” which are used in the same way is conducted. And the others are excluded from the target.

By using this cleaned data and “Defective system”, analyzing the relational combinations of three categories of defect factors is conducted. For analyzing, correspondence analysis, co-occurrence network and self-organizing map are applied to two relationships between systems and parts and between parts and primary factors. In the following, results of analysis the data which “More (leaking)” is included and not “GASU (gas)” are not included as a result of liquid leakage.

#### 4.1 Deriving relative combinations of defect factors by correspondence analysis

By applying correspondence analysis, deriving relative combinations between systems and parts and between parts and primary factors. Condition of each words are set as follows.

- **System:** 11 types from the classification of defect systems which is used at the reports of Japanese Ministry of Land, Infrastructure and Transport
- **Parts:** General nouns from morphological analysis by KH Coder
- **Primary factor:** 16 types (position, strength, shape, structure, material, design, evaluation, error, processing, control, work instruction method, tightening, mounting, heat treatment, welding, other manufacturing) from investigation
In addition, condition for depiction is set as the top 60 “words with remarkable differences” (words with large variation in appearance ratio among qualitative variables) in KH Coder, and condition for extraction is set as nearby 5 parts for each systems from a result of systems and parts and nearby 1 primary factors from a result of parts and primary factors. With these conditions, the map of systems and parts and the map of parts and primary factors which show 1st and 2nd principal component scores are derived. As an example, relations are obtained such as relationship between parts with fuel system from the map.

4.2 Deriving the relative combinations between parts and primary factor by co-occurrence network
By using the method of co-occurrence network, depict co-occurrence network diagram for parts and primary factors are conducted. By using this, finding the strong combinations of co-occurrence relationships is conducted. As the condition for words, parts are set as same as correspondence analysis and primary factors are set as the noun of the irregular conjugation of the line of “Sa” from morphological analysis. For other conditions, Jaccard coefficient is used as a calculation method of co-occurrence, and 0.2 or over is limit of Jaccard coefficient for depicting. With these conditions, the map of co-occurrence network is derived.

4.3 Deriving relative combinations between parts and primary factor by self-organization map
By using the algorithm of self-organization map, extracting nearest word of primary factors for each words of parts as strong combinations are conducted. For analyzing, conditions for each meaning are as same as the analysis of co-occurrence network. In addition, each condition of self-organizing map is set as follows.

- Number of nodes: 400
- Number of clusters: 8
- Number of ordering phase: 1000 times
- Number of convergence phase: 200,000 times

With these conditions, the map is derived as the result of analysis. The results shows some parts, that are closing to the word of others. When condition of extraction is set as nearby 5 parts for each system from a result of systems and parts and nearby one primary factor from a result of parts and primary factors, Table. 4 is shown as the results from the above analysis as derived combinations of factors in fuel system which cause liquid leakage.

| systems      | parts                  | primary factors            |
|--------------|------------------------|----------------------------|
|              | Correspondence analysis| Correspondence analysis    | Co-occurrence network     | Self-organization map |
| Fuel system  | tank                   | Strength                   | mounting                  | airtight               |
|              | engine                 | Work instruction method    | mounting                  |                       |
|              | fuel                   | position                   | molding                   |                       |
|              | gas                    | evaluation                 | airtight                  |                       |
|              | bellows                | Position                   |                           |                       |

In conclusion, it can be said that this extracting way can derive 13 primary factors of 5 parts in fuel system as the factors of liquid leakage.
4.4 Discussion
In order to check the amount of information of Table 4, aggregating the number of each pairs from the recall information which is related to liquid leakage is conducted. First, result of system and parts is shown in Table 5.

Table 5. Result of aggregating information which refers the word of systems and parts in Table 4

| systems word | number | parts word | number |
|--------------|--------|------------|--------|
| Fuel system  | 85     | tank       | 23     |
|              |        | engine     | 12     |
|              |        | fuel       | 59     |
|              |        | gas        | 3      |
|              |        | bellows    | 0      |

Table 6. Result of aggregating primary factors of “tank” which is referred in Table 3

| parts word | Primary factors from Correspondence analysis | Primary factors from Co-occurrence network | Primary factors from Self organization map |
|------------|---------------------------------------------|-------------------------------------------|-------------------------------------------|
| tank       | Strength                                   | mounting                                  |                                           |
|            | 23                                         | 4                                         | 1                                         |

Table 7. Result of aggregating primary factors of “tank” which is not referred in Table 6

| Primary factors       | number |
|-----------------------|--------|
| Tightening            | 5      |
| Work instruction method | 3      |
| material              | 2      |
| welding               | 2      |
| design                | 1      |
| position              | 1      |
| shape                 | 1      |
| structure             | 1      |
| Other manufacturing   | 2      |

From Table 5, it can be found that there are some words which score a few numbers like “bellows” and words which do not mean parts like “engine”. From here, it can be said that in appropriative words are existed in this result. It would be considered that there are 3 factors that cause it. First factor is non preciseness regarding the general noun as the word of parts. Second factor is contaminating unnecessary information (ex. fuel filter” attached to the engine” is ...). And third factor is the difficultness of deriving the relational pair by correspondence analysis when there the parts is related multiple systems.

Based on this discussion, aggregating the numbers of primary factors for “tank” which is most likely to mean parts is conducted. The result of primary factors which is referred in Table 4 is shown in Table 6, and the result of others are shown in Table 7. From comparing with these 2 results, it can be found that the number of “strength” which derived as primary factors related to “tank” is not the biggest number. From here, it can be said that there are cases that cannot extract most expectable primary factors in this way of extraction. It is considered that there
are 2 factors that cause it. First is the less of the number of primary factors extracted for each method. And second is the difficulty of the deriving the pairs that is not unique by relevance.

5. Revealing the factors of important defect phenomena by integrating the results

There are 2 steps for integrating the results in proposal way. First step is selecting step which select the important defect phenomenon by 3 conditions of scores from the analysis of the vehicle defect information. Last step is revealing step which analyze the recall information which is relative to the important defect phenomenon from selecting step. By through these steps, as a result, the factors of important defect phenomena are derived.

5.1 Select the important defect phenomena

From the analysis of the vehicle defect information, 3 scores which is considered to show the numbers, importance and effect of defect phenomenon are calculated. From here, the dependency pairs which all of 3 scores get bigger number defined the important defect phenomenon. In this research, conditions of each scores are set as follows.

- The number of occurrences of the pairs: 30 or more
- BM 25 score of dependency pairs: 10.0 or more
- the BM 25 score of modified words: 10.0 or more

With these conditions, 13 dependency pairs are extracted as the important defect phenomenon from 15961 dependency pairs. These pairs are shown in Table.8.

5.2 Revealing the factors of selected defect phenomena

In the analysis of the recall information, combinations of defect factor are derived by 3 methods. In this step, by analyzing the recall information which is related to the important defect phenomenon from selecting step, deriving the factors of important defect phenomenon is conducted. When analyze for liquid leakage as one of the important defect phenomena as an example, reveal 97 primary causes in 49 parts of 10 devices (no overwrap) are derived. It is considered that this result has 2 specific point. 1st point is based on priority orientate. 2nd point is showing the categories of defect factors systematically. It is considered that these points are useful for kaizen works.

6. Conclusion

In this research, analyzing both information of vehicle defect is conducted. At the analysis of the vehicle defect information, 3 scores which is considered to show the numbers, importance and effect of defect factors for dependency pairs are indicated. At the analysis of the recall information, even if there are problems of accuracy, relative combinations of defect factors are shown. Finally, as a result of integrating analysis by proposal way, the factors of important defect phenomenon which has specific point for kaizen works are derived.

However, it should be careful that this result derives only the factors that is happened in the past, cannot reveal the factors which is not happened before. Therefore, it is considered that this method cannot reveal the factors proper for the products that changes the design significantly. On the other hand, since these are factors that has been recalled, we consider that these are factors that are difficult to find in ordinary design and preventive activities.

Acknowledgement

The authors sincerely thank the reviewers for their constructive comments.
Integration of defect and recall information, Yamamoto et al.

References:

Kitazawa, K. and Osada, H. (2008), “Analysis Method of Quality Accidents by Using Public Information and its Result-Analysis of Recall Data in Automobile Industry”, *Journal of the Japanese society for quality control*, Vol.38, No.1, pp.147-155

Misue, K. and Watanabe, I. (1999), “Visualizing techniques of associative relationship for text mining”, *IPSJ SIG technical reports*, 1999-Fl-055, Vol.1999, No.57, pp.65-72

Nishiura, Y. and Yamada, S. (2010), “An Approach for Exploring Design Review Items Based on Product Problem Information”, *Journal of the Japanese society for quality control*, Vol.40, No.4, pp.411-419

Ohsawa, Y., Nels E. Benson, and Yachida, M. (1999), “KeyGraph: Automatic Indexing by Segmenting and Unifying Co-occurrence Graphs”, *IEICE TRANSACTIONS on Information and Systems*, Vol. J82-D1, No.2, pp.391-399

Takigawa, M. and Yamana, H. (2016), “Proposal of Word Importance Calculation Method for Specific Fields and Adaptation to Expertise Estimation on Twitter”, *Proceeding FIT : forum on information technology 2016*, Vol.15, No.2, pp.1-7

Type Approval and Recall Division, Road Transport Bureau, Ministry of Land, Infrastructure, Transport and Tourism. (2018), “The analysis result about the content of the vehicle recall of the 29th year of the Heisei period”, retrieved 10th December 2018, from [http://www.mlit.go.jp/jidosha/carinf/rel/common/data/h29recallbunseki.pdf](http://www.mlit.go.jp/jidosha/carinf/rel/common/data/h29recallbunseki.pdf)

Type Approval and Recall Division, Road Transport Bureau, Ministry of Land, Infrastructure, Transport and Tourism. (2018), “The result of aggregation of the vehicle defect information of the 4th quarter of the 29th year of the Heisei period”, retrieved 10th December 2018, from [http://www.mlit.go.jp/jidosha/carinf/rel/common/data/quarter_29_4.pdf](http://www.mlit.go.jp/jidosha/carinf/rel/common/data/quarter_29_4.pdf)

Type Approval and Recall Division, Road Transport Bureau, Ministry of Land, Infrastructure, Transport and Tourism. (2018), “Searching vehicle defect information”, retrieved 10th December 2018, from [http://carinf.mlit.go.jp/jidosha/carinf/opn/index.html](http://carinf.mlit.go.jp/jidosha/carinf/opn/index.html)

Type Approval and Recall Division, Road Transport Bureau, Ministry of Land, Infrastructure, Transport and Tourism. (2018), “Searching recall information”, retrieved 10th December 2018, from [http://carinf.mlit.go.jp/jidosha/carinf/ris/index.html](http://carinf.mlit.go.jp/jidosha/carinf/ris/index.html)

Yoshioka, K. and Koeda, M. (2012), “Related word extraction and word analysis using BM 25”, *Proceedings of 74th National Convention of IPSJ*, pp.553-554

Author’s biographical notes

**Mr. Yasuyuki Yamamoto** is a master student, Center for Open Systems Management, School of Science for Open and Environmental Systems, Keio University. He receives his Bachelor of Science Degree from Keio University.

**Dr. Shu Yamada** is a professor, Faculty of Science and Technology, Keio University. He teaches applied statistics and quality management at undergraduate and graduate schools. His research interests focus on total quality management, design of experiments for quality, statistical quality control and so forth.