Journal of Intelligent Manufacturing (2020) 31:249–265
https://doi.org/10.1007/s10845-019-01466-z

A data-driven approach for constructing the component-failure mode matrix for FMEA

Zhaoguang Xu[1] · Yanzhong Dang[1] · Peter Munro[2] · Yuhang Wang[1]

Received: 24 August 2018 / Accepted: 10 February 2019 / Published online: 18 February 2019

Abstract
Failure mode and effects analysis (FMEA) is one of the typical structured, systematic and proactive approaches for product or system failure analysis. A critical step in FMEA is identifying potential failure modes for product sub-systems, components, and processes, for which component-failure mode (CF) knowledge is necessarily needed as an important source of knowledge. However, this knowledge is usually acquired manually based on historical documents such as bills of material and failure analysis reports, which is a labor-intensive and time-consuming task, incurring inefficiency and plenty of mistakes. Nevertheless, few existing studies have developed an effective and intelligent approach to acquiring accurate CF knowledge automatically. To fill the gap, this paper proposes a method to construct the CF matrix automatically by mining unstructured and short quality problem texts and mapping as well as representing them as CF knowledge. Starting with mining the frequent itemsets of failure modes through Apriori algorithm, the method uses the semantic dictionary WordNet to find synonyms in the set of failure modes, based on which the standard set of failure modes is finally built. Subsequently, upon the previous work and components set, we design the component-failure mode matrix mining (CFMM) algorithm and apply it to establish the CF matrix from unstructured quality problem texts. Lastly, we examine the quality data of the seat module of an automobile company as a case study in order to validate the proposed method. The result shows that the failure mode extraction method with standardized features can extract failure modes more effectively than the FP-growth and K-means clustering methods. Meanwhile, the devised CFMM algorithm can extract more combinations of CF than the FP-growth method and build a richer CF matrix. Although different industries have distinct domain characteristics, our proposed method can be applicable not only to manufacturing but also to other fields needing FMEA to enhance product and system reliability.

Keywords Failure mode and effects analysis · Component-failure mode matrix · Data mining · System reliability · Automotive industry

Introduction

Failure mode and effects analysis (FMEA) is a systematic activity for revealing potential faults when a firm does the planning of developing a product or new production methods, and for implementing appropriate actions to avoid faults, which ultimately improves product quality and reliability (Stamatis 1995; Liu et al. 2016). By definition, failure mode refers to the termination of the ability of a system to perform a required function or its inability to perform within previously specified limits (ISO/IEC-15026-1 2013) and includes both known and/or potential failures and problems that may incur customers' dissatisfaction and poor evaluation, and thus endanger the reputation of the entire organization (Asan and Soyer 2016). In practice, FMEAs come in various forms, such as Design FMEA (DFMEA), Process FMEA (PFMEA) and System FMEA (Pfeifer 2002) according to different emphasis and objectives. When creating a DFMEA, listing all potential failure modes for each component is a very important step (Brook 2006) that is critical for creating failure-free designs (Arunajadai et al. 2004). While anticipating every failure mode is impossible, the development team should formulate a list of potential failure modes as extensive as possible (Goel and Graves 2007). Most firms usually manually analyze, collate and summarize historical

✉ Zhaoguang Xu
zhaoguang0920@126.com

1 Institute of System Engineering, Dalian University of Technology, Dalian 116024, China

2 BMW Brilliance Automotive Ltd., Shenyang 110143, China

documents such as the product function decomposition model, bill of materials (BOM) and failure analysis report to obtain the failure modes of components and obtain the corresponding relationship between the components and the failure modes (Wang et al. 2011).

DFMEA is a commonly used but significant tool in product design and development to take full account of the problems involved in the process of production, transportation, and use of products, to bring all possible problems into the scope of prevention, and to do a good job of preventive measures and solutions in advance. The creation of DFEMA first needs to know which failure modes have occurred in the product components. Component-failure mode matrix is an important source of knowledge in this process.

However, there are some assignable drawbacks of manually acquiring failure modes and their associations with components under a DFMEA. First, the source of failure modes knowledge is very fragmented. When these documents are missing or difficult to find, the component-failure mode (CF) knowledge will be incomplete. Second, a large number of failure-mode types make it difficult for an enterprise to build a firm-level failure-mode library. Instead, different departments in a company usually use their own description vocabulary when describing the same failure mode or the same description for two marginally different failures (Tumer et al. 2003). Also, manually building CF knowledge is a time-consuming and labor-intensive activity. On the one hand, the designers who create DFMEA are far away from the production process, lack of understanding of the product quality problems that may occur in the production process; and the data of product quality problems scattered in the production process form an information isolated island, which is difficult for designers to use. On the other hand, the employees who record product quality problems often adopt according to their own habits when describing the same problem. Different words are used to describe the failure mode, which results in the ambiguity of the designer's perception of failure mode.

In response to these problems, many scholars have studied the extraction of failure modes (Collins et al. 1976; Arunajadai et al. 2002; Tumer et al. 2003; Wani and Jan 2006; Chen and Nayak 2007; Wijayasekara et al. 2014; Chang et al. 2015; Rajpathak and De 2016; Kai et al. 2015; James et al. 2017; Meng et al. 2017) by utilizing classification, clustering and other methods to extract failure modes. Among these studies, few studies focus on the standardization of failure modes. Turner et al. (2003) provide a standard failure-mode taxonomy with a definition of three levels for each failure mode by analyzing operational failure reports from a problem and failure reporting database at Jet Propulsion Laboratory. However, these failure modes were based on predetermined failures (Roberts et al. 2003), and the authors did not mention the way how to build up the standard failure mode taxonomy. Beksinska et al. (2007) developed a standardized list of terms and definitions of failure modes for female condoms. However, there were only eight kinds of failure modes, and all of them were compiled by members of the WHO Technical Review Committee which is not suitable for complex products and systems. Other scholars have focused on functional-failure mode (EF) matrix construction (Arunajadai et al. 2002; Tumer and Stone 2003; Chang et al. 2015) and fault dependency matrix (D-matrix) mining (Singh et al. 2010; Rajpathak and Singh 2014; Deore 2015; Thombare and Dole 2015; Jenifa and Balachander 2015; Mendhe and Hande 2017). These studies presupposed that the CF matrix is known, whereas few studies revolve around mining the CF matrix from a large number of texts.

However, in the production process, as well as quality management activities, a large amount of data related to the quality problem are generated and accumulated. In our viewpoint, the knowledge embedded in the unstructured quality problem data provides insight into component failure. Meanwhile, text mining is important because it can automatically discover knowledge assets hidden in the unstructured text (Hearst 1999; Khilwani and Harding 2016). Therefore, using a text mining method to automatically mine text instead of manually acquiring the failure modes of product components and the relationship between components and failure modes from a large number of quality problem data can make it possible to predict each failure mode. The output of this method will provide a data foundation for DFMEA, improve the efficiency of building a DFMEA knowledge base, and improve product and system reliability and customer satisfaction.

Stated thus, there is a paradox: on the one hand, we want the description of employees to be uniform and accurate, such as building a standard failure-mode set manual beforehand but at the expense of the authenticity of the information. After all, we cannot exhaust all failure modes and ensure that they fully match the actual situations staff encounter. On the other hand, employee's personalized description guarantees the authenticity and accuracy of the information records, but the description is probably confusing, repetitive and logically poor, which brings great challenges to failure patterns recognition and CF matrix construction. Unfortunately, the existing methods have fairly limited ability to solve this paradox.

To address the previous challenge, this paper solves the issues of standardization of failure modes and automatic construction of a CF matrix based on data with implicit failure modes. In this study, the WordNet-based text mining method was initially built to create a standard failure mode library. Then, this paper proposes a component-failure mode matrix mining (CFMM) algorithm. Based on the standard failure modes constructed above and the existing product components, the algorithm extracts the CF matrix from the quality problem text, which can be treated as the component failure mode knowledge and the basic knowledge of FMEA. In particular, this paper makes the following contributions:

1. Different from the existing methods of failure-mode extraction, this paper considers the phenomenon that different departments use a different vocabulary to describe the same problem or failure mode and further studies the problem of failure mode standardization. Based on the semantic dictionary WordNet, this paper recognizes and unifies synonymous failure modes and then constructs a standard set of them, which provides a common vocabulary across business units, providing a more comprehensive and standard knowledge resource for DFMEA.
2. Based on historical data of components and failure modes, the CFMM method is employed to build a CF matrix automatically, which is more efficient and reliable than traditional experience-based, and brainstorm-based approaches. It can also be combined with traditional approaches to building the CF matrix better. The CFMM algorithm in this paper covers both significant and insignificant FM in frequent itemsets more comprehensively and constructs the standard failure mode and the correlation matrix between components more completely with higher accuracy.
3. In the existing research on failure modes, the data source of most texts is after-sales data (Rajpathak and De 2016), maintenance text data (Chen and Nayak 2007), existing FMEA, FMECA and other data. To the best of our knowledge, this paper is the first to use the product quality problem-solving data in the manufacturing process as the data source for failure mode extraction. Note that our proposed method can also be compatible with data-driven FMEA construction based on sales data.

The remainder of this paper is organized as follows. In the next section, we give a literature review to related studies. Then, in "Research framework" section, the general research framework is provided to help guide the reader through the steps involved. In "Standard failure mode set construction" section, the WordNet-based method is introduced to build up the standard failure mode set. "Component-failure mode matrix mining" section presents the CFMM algorithm. "Case study" section takes the seat module in an automotive company as an example to perform our methodology and conduct relevance analysis. Then, we conclude our paper by discussing the benefit of our methodology and summarizing the main findings in "Conclusions and future work" section.

Literature review

Failure mode acquisition

Many text mining-related methods have been used by scholars to obtain failure modes. Some of them have studied the methods of failure mode classification. Based on the failure-experience matrix (Collins et al. 1976), Arunajadai et al. (2002) classified incremental bills of materials with recorded failure information into corresponding default failure modes. Wijayasekara et al. (2014) divided the software code failure data into hidden impact bugs and regular bugs. Wu et al. (2017) proposed a classification tree kernel-based support vector machine to identify bearing failures. However, the classification method presupposed the type of failure mode and usually set a few categories of failure modes. This approach is not applicable to complex systems, such as automotive products, which may have many failure modes.

In addition, many scholars have studied how to use the clustering method to acquire failure modes. Based on the artificially established failure modes and their frequency, Wani and Jan (2006) adopted the "K clustering" method to determine the failure mode group in the conceptual design phase of the mechanical system. Chen and Nayak (2007) studied the method of automatically extracting failure modes from maintenance text datasets using Ward's agglomerative method and the similar histogram clustering method. Arunajadai et al. (2004) constructed a similarity matrix between failure modes and then obtained failure mode groups through hierarchical clustering.

Moreover, some scholars clustered failure modes for better analysis based on historical data of FMEA. Chang et al. (2015) clustered the failure modes in the FMEA and converted the failure modes in complex FMEA worksheets into a tree structure by constructing the ETree learning algorithm. Kai et al. (2015) studied the Euclidean distance-based similarity measure and fuzzy adaptive resonance theory neural network for the similarity analysis and clustering of failure modes in FMEA. Meng et al. (2017) performed K-means clustering on the preprocessed software failure text and selected representative failure texts from the clusters as cluster labels.

Apart from the methods of classification and clustering, some scholars have studied the ontology method to extract failure modes. Rajpathak and De (2016) provided an ontology-based approach for identifying failure modes from repairing verbatim data. James et al. (2017) studied the construction of failure knowledge ontology based on historical data on maintenance and services.

Among the above methods, few studies have considered the case where the description of the failure mode is not uniform. However, quality, maintenance or after repair records are usually recorded by different people or departments in different ways. One of the most limiting aspects of FMEA is the lack of a standard vocabulary to describe functionality and failure modes accurately and without ambiguity (Schneider 2003). Although some studies have emphasized that standardization of failure mode vocabulary help to effectively maintain and utilize the knowledge base and provides a standard failure mode vocabulary (Arunajadai et al. 2002), none of these studies provided a systematic approach to standardizing failure modes.

Considering the limitations of existing literature, this paper proposes a text mining method based on WordNet to construct a standard failure-mode set. Compared with previous studies, this method applies to a large number of failure-mode sets and can extract and construct standard failure-mode sets without consuming considerable labor and time.

Component-failure mode matrix

CF knowledge is a representation of the potential failure modes of product subsystems and components in the FMEA and can be represented by a $n \times m$ CF matrix (Tumer and Stone 2003; Arunajadai et al. 2004), where m is the total number of failure modes occurring across all n components. In this matrix, a '1' is placed for a component in the cell corresponding to the failure mode the component experienced and a '0' is placed in the other cells (Arunajadai et al. 2002). Except for the binary information of failure modes for a given component (Xu et al. 2018), the likelihood or frequency of occurrence data can also be encoded in CF (Wang et al. 2011).

Many studies involving component and failure modes used the CF matrix as a known resource to study other issues, such as functional-failure mode (EF) matrix construction and fault dependency matrix (D-Matrix) construction. The EF matrix relates the failure modes to the elemental functions. Each element in the matrix indicates whether any component solving function has ever failed by a failure mode (Arunajadai et al. 2002; Tumer and Stone 2003; Chang et al. 2015). Most studies centered on the construction of the EF matrix by formula $EF = EC \times CF$, where EC represented a functional component matrix. Another area of study related to component-failure mode considered D-matrix. Unlike the CF matrix, the D-matrix indicates the dependencies between observable symptoms and failure modes. It is a system diagnostic model for capturing hierarchical system-level fault diagnosis information (Rajpathak and Singh 2014). In the D-matrix matrix, rows represent combinations of components and failure modes, and columns represent symbols. Singh et al. (2010) introduced three types of D-matrices and introduced sources of D-matrices, such as historical field fault data, engineering schematics, and failure modes, and impact and critical analysis (FMECA) data. Based on the construction of fault diagnosis ontology, Rajpathak and Singh (2014) applied ontology-based text mining algorithms to identify necessary artifacts, such as parts, symptoms, failure modes and their dependencies, from unstructured repair verbatim texts in the automotive field. Based on this research, many scholars have conducted similar research. Deore (2015) and Thombare and Dole (2015) described an ontology-based text mining method for automatically building and updating the D-matrix by mining thousands of repaired verbatim data collected during diagnostic events. Mendhe and Hande (2017) further studied the representation of the D-matrix in a graph. Jenifa and Balachander (2015) introduced a method to construct the D-matrix with the help of the FP growth algorithm such that the best-practice repair actions can be discovered.

However, all of these studies assumed that the CF matrix or relationship between components and failure modes is known (Liu et al. 2017), and in the case studies of these papers, the number of components and failure modes were relatively small. Thus, the CF matrix could be provided based on experience. Unfortunately, complex products or systems contain many components, and there are many failure modes, which makes it time-consuming and labor-intensive to acquire the CF matrix manually. Therefore, it poses significant challenges to obtaining the component and the failure mode relationship automatically. Given such constraints, this paper provides a CFMM method based on historical text data to automatically construct a CF matrix, which fills the gap in the field.

Research framework

The goal of this study is to automatically obtain the CF matrix from a large number of unstructured quality problem data. For exposition clarity, we illustrate the research framework and the process of CF matrix extraction in Fig. 1. The raw data are first preprocessed; then, a standardized failure-mode set is constructed by the following steps, including failure mode frequent itemset mining and frequent itemset standardization. Moreover, the nonstandard failure mode text in the existing problem title set is replaced with the standard failure mode to form a new problem title set. Furthermore, the CF matrix mining algorithm is designed. Based on the standard failure mode set and the existing component set, the algorithm is used to extract the CF matrix from the processed quality problem text. The part covered by the red line in Fig. 1 is the focus of this paper. In "Standard failure mode set construction" section, the process and method of constructing the standard failure mode set are introduced in detail. The "Component-failure mode matrix mining" section serves to describe the CFMM algorithm in detail.

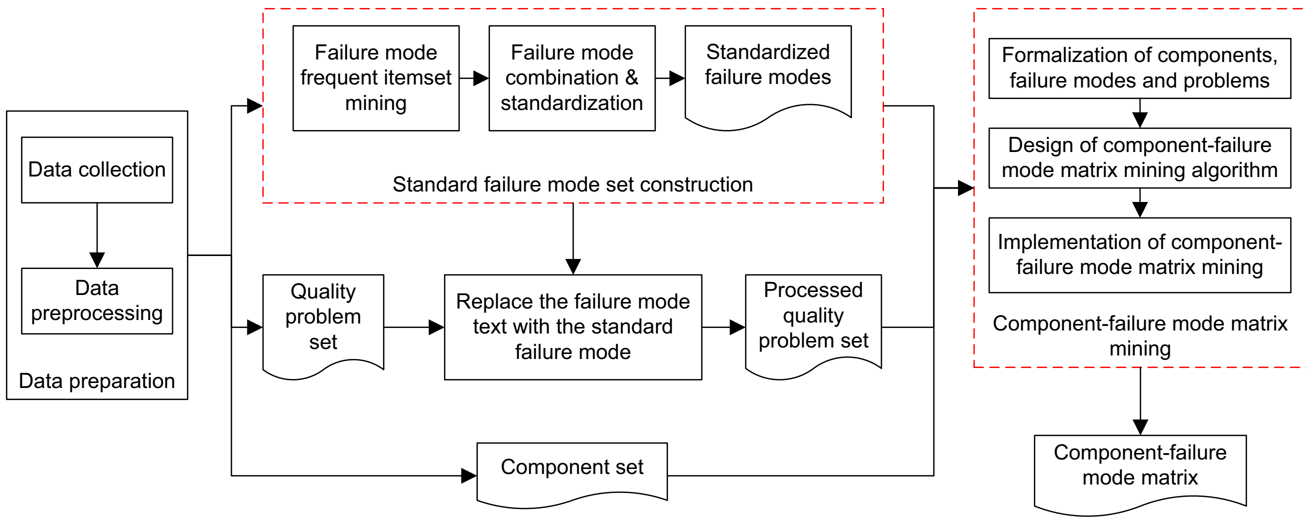

Fig. 1 The research framework

Standard failure mode set construction

According to the research framework in Fig. 1, we can see that the method of constructing a standard failure mode set includes the following steps: text preprocessing, failure mode frequent itemset mining, and failure mode standardization. Text preprocessing operations include removing stopwords, converting abbreviations into complete words, and removing context words. Failure mode frequent itemset mining includes part-of-speech tagging, extracting frequent itemsets using the Apriori algorithm, and pruning. Failure mode standardization operations include WordNet-based failure mode synonym identification and failure mode combination and standardization.

This paper collects related problem corpora from the database of quality problem-solving. $T = \{t_1, t_2, \ldots, t_l\}$ is the original quality problem text set, where t_s represents the s_{th} problem, $s = 1, 2, \ldots, l$. First, the original text should be preprocessed by deleting stopwords, translating acronyms into complete words, and deleting context words. In the step of removing stopwords, the stopwords in the text are deleted based on the current stopwords. At the same time, a dictionary of domain abbreviations is built to convert abbreviations in the original text into complete words. Many scholars have proposed different methods to address the abbreviation ambiguity problem (Wu et al. 2015; Kim and Yoon 2015). However, the abbreviations have domain characteristics, and the quality problems are titled in short texts. These abbreviations usually have a unique sense. Therefore, after identifying the abbreviation in the original text, it is replaced with the word in the abbreviation database, and the disambiguation processing is not performed. In addition, the original text contains information such as components, which contain words that are different from the words contained in the failure mode. Therefore, to focus on the failure mode, so as not to interfere with this information during the failure mode extraction process, words contained in the components in the original text are deleted according to the known component set $C = \{c_1, c_2, \ldots, c_n\}$.

Then, in the failure mode frequent itemsets mining step, all the problem texts are part-of-speech tagged, and all the word sets $U = \{u_1, u_2, \ldots, u_o\}$ in the preprocessed document are obtained. Moreover, these words are used as items, and each problem record is taken as a transaction unit to create associated transaction data. According to the Apriori algorithm (Han et al. 2011), itemsets satisfying the minimum support threshold α are extracted from associated transaction data as candidate itemsets. The set of candidate itemsets is represented as $F_1' = \{f_1', f_2', \ldots, f_j', \ldots f_v'\}$, in which f_j' is the j_{th} frequent itemset. Each candidate itemset can be seen as a failure mode. According to the experience of domain experts, a failure mode contains no more than three words, so this paper does not consider frequent itemsets with more than three items. Moreover, frequent itemsets do not belong to failure modes, such as road, cobblestone and other words that indicate the situation of the problem, and the words that indicate a location such as right, left, rear, front, and middle. Therefore, combined with expert experience, this paper performs artificial pruning on candidate itemsets and then obtains a new set of candidate itemsets $F_2' = \{f_1', f_2', \ldots, f_j', \ldots f_\delta'\}$, where $\delta < v$.

The new set of pruned frequent itemsets contains a large number of failure modes. If the result is used as a failure mode set, there will be frequent itemsets that may represent the same failure mode. Therefore, it is necessary to standardize the set of candidate itemsets F_2' and merge the different frequent itemsets that are described by synonyms. In quality management, developing a generic description vocabulary that can be understood by the various departments is a challenge for the organization. It would take many workforces and material resources to identify synonyms or phrases artificially from hundreds of words or phrases manually. Therefore, some methods are often used to identify synonyms, such as identifying synonyms in dictionary annotations (Blondel and Senellart 2002; Muller et al. 2006; Wang and Hirst 2012), vocabulary cooccurrence algorithms in large corpora (Baroni and Bisi 2004), and search engine-based methods for identifying synonyms in web (Yates and Etzioni 2009; Cheng et al. 2012). However, it is difficult to identify the synonymous failure modes by these methods since the quality problem text is usually short and contains limited information. Of course, there may be a phenomenon of polysemy for the corpora on the Internet, whereas the failure mode in this research is often a non-polysemy noun which has strong domain characteristics. Therefore, each frequent itemset with a support count can be considered a failure mode with only a single meaning. In the step of standardization of failure modes, this paper introduces a synonym extraction method based on WordNet. WordNet is a cognitive linguistics-based English dictionary designed by psychologists, linguists and computer engineers at Princeton University. It organizes vocabulary information based on word meaning rather than word form. WordNet groups them according to the meaning of the terms. Each group of words with the same meaning is called a Synset (Fellbaum 2000). WordNet is used to find synonymous relationships between all words in the set of candidate itemsets and build a synonym set. For each set of synonymous frequent itemsets, the frequent itemsets with the highest support count are taken as the standard failure mode of the group. Finally, combined with the experience and opinions of domain experts, the results are revised, and a standard set of failure modes is constructed. For example, according to WordNet, f_i' and f_j' are synonymous, and the support count of a frequent itemset f_j' is higher than f_i', then these two failure modes are unified into f_j' as the standard failure mode, and the new support count of f_j' will be the sum of first support count of f_j' and f_i'. According to this rule, this paper constructs a synonymous failure mode set and obtains a new set of frequent itemsets $F = \{f_1, f_2, \dots, f_m\}$ as a standard failure mode set, in which $m < \delta$.

Component-failure mode matrix mining

Notation and formalization

We use the titles of the quality problems as a link to construct the CF matrix of the existing component set and the standard failure-mode set. Here we give the notations and definitions which will be used in the CFMM algorithm.

Automotive components are an essential part of a car. They usually consist of multiple parts and have specific functions. The component set can be formalized as

$$C = \{c_1, c_2, \dots, c_n\},$$

where c_i indicates the i_{th} component, $i = 1, 2, \dots, n$.

Failure mode refers to the termination of the ability of a system to perform a required function or its inability to perform within previously specified limits. It is the result of the failure mechanism (cause of the failure mode). For example; a fully fractured axle, a deformed axle or a fully open or fully closed electrical contact are each a separate failure mode of a DFMEA. The failure mode set can be formalized as

$$F = \{f_1, f_2, \dots, f_m\},$$

where f_j represents the j_{th} failure mode, $j = 1, 2, \dots, m$.

A problem title is a comprehensive refinement of the problem, usually recorded in the form of short text, and contains information about components and failure modes. Problem title set can be represented by

$$T = \{t_1, t_2, \dots, t_l\},$$

where t_s represents the s_{th} problem title, $s = 1, 2, \dots, l$.

The text of components, failure modes, and problem titles can only be computed after mathematical representation. This paper uses the bag of words (BOW) model to represent these texts. In information retrieval, the BOW model assumes that for a document, its word order and grammar, syntax and other elements are ignored, and it is only regarded as a collection of several words. The appearance of each word in the document is independent and does not depend on whether other words appear. Before using the BOW model to represent text, it is necessary to create a dictionary. A dictionary consisting of the words contained in the text of the component, failure mode, and problem title can be represented as

$$W = [w_1, w_2, \dots, w_e],$$

where w_τ represents the τ_{th} word in the dictionary, $\tau = 1, 2, \dots, e$.

Based on the dictionary, this paper uses BOW to represent the component set as the document-term matrix. The document-term matrix of components can be formalized as

$$CW_{ne} = \left(cw_{i\tau}\right)_{n\times e} = \begin{pmatrix} cw_{11} & cw_{12} & \cdots & cw_{1e} \\ cw_{21} & cw_{22} & \cdots & cw_{2e} \\ \vdots & \vdots & \ddots & \vdots \\ cw_{n1} & cw_{n2} & \cdots & cw_{ne} \end{pmatrix},$$

where $cw_{i\tau} = 1$ indicates that component c_i contains the word w_τ, and $cw_{i\tau} = 0$ indicates the opposite.

Similarly, the document-term matrix of failure modes can be formalized as

$$FW_{me} = \left(fw_{j\tau}\right)_{m\times e} = \begin{pmatrix} fw_{11} & fw_{12} & \cdots & fw_{1e} \\ fw_{21} & fw_{22} & \cdots & fw_{2e} \\ \vdots & \vdots & \ddots & \vdots \\ fw_{m1} & fw_{m2} & \cdots & fw_{me} \end{pmatrix},$$

where $fw_{j\tau} = 1$ indicates that failure mode f_j contains the word w_τ, $fw_{j\tau} = 0$ indicates the opposite.

The document-term matrix of problem titles can be formalized as

$$TW_{le} = \left(tw_{s\tau}\right)_{l\times e} = \begin{pmatrix} tw_{11} & tw_{12} & \cdots & tw_{1e} \\ tw_{21} & tw_{22} & \cdots & tw_{2e} \\ \vdots & \vdots & \ddots & \vdots \\ tw_{l1} & tw_{l2} & \cdots & tw_{le} \end{pmatrix},$$

where $tw_{s\tau} = 1$ indicates that problem title t_s contains the word w_τ, $tw_{s\tau} = 0$ indicates the opposite.

CF matrix is an $m \times n$-dimensional matrix, it can be represented as

$$CF_{nm} = \left(cf_{ij}\right)_{n\times m} = \begin{pmatrix} cf_{11} & cf_{12} & \cdots & cf_{1m} \\ cf_{21} & cf_{22} & \cdots & cf_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ cf_{n1} & cf_{n2} & \cdots & cf_{nm} \end{pmatrix},$$

where m is the total number of failure modes occurring across all n components, and cf_{ij} indicates the number of times that component c_i has experienced failure mode f_j.

Assumptions and algorithm

In this section, we provide a novel text mining method for mining the relationships of components and failure modes. For the ease of quality exposition, we need to give out some assumptions as follows.

Assumption 1 Each title only contains one component.

Assumption 2 Each title only contains one failure mode.

In fact, a problem title in practice usually contains the following information, including the problem situation, the components in which the problem occurred, and what the problem is. People who input problems into the system will not describe the problematic component and the failure mode multiple times in the title. These titles form a concise representation of the most important message of a document (Mangnoesing et al. 2012), and they are often concise and rarely have complex statement expressions (Miao et al. 2008). Moreover, the method also has some reference for extracting related information from the titles described by short text in other scenarios. However, for some non-title content, for example, an article or lengthy text, there may be different methods for different research needs.

In this method, we first need to find out which components and failure modes each title contains. Through the formula below, we can obtain the association between the title and the component.

$$TC = \left(tc_{si}\right)_{l\times n} = TW \times CW^T = \begin{pmatrix} tc_{11} & tc_{12} & \cdots & tc_{1n} \\ tc_{21} & tc_{22} & \cdots & tc_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ tc_{l1} & tc_{l2} & \cdots & tc_{\ln} \end{pmatrix} \tag{1}$$

In formula (1), CW^T represents the transposition of the document-term matrix of components. In each row of the matrix TC, the result of the multiplication is a number indicating the number of identical words in the component and the problem title. For example, tc_{11} indicates the number of identical words in the first component and the first problem title. For each title, we need to find a component in all components so that the value of this multiplication is the largest, which indicates that the number of identical words in this component and this problem title is the highest. This component is the component that the title contains. For example, if $\max\left(tc_{11}, tc_{12}, \ldots, tc_{1n}\right) = tc_{12}$, then title 1 contains component 2. If a problem title can correspond to multiple components, we take the component that contains the least number of words as the component corresponding to the problem title. For example, the problem title is "Seat belt noise," there are two components in the component set that are "seat belt" and "seat belt buckle." Then, the dictionary will be {seat, belt, noise, buckle}. According to the above method, the results of multiplying the seat belt and the seat belt buckle by the problem title are both 2. However, the corresponding component in the title should be the "seat belt," not the "seat belt buckle."

After this step, we get the correspondence between all titles and components and store the subscripts of the components identified in all titles in an array, which then generates a collection of components corresponding to all the problem titles. Moreover, the occurrence frequency of each component is calculated as the statistical result of the component identified from all the titles.

As with the above method of obtaining the association between the titles and the components, we can get the association between the titles and the failure modes by the following formula.

$$TF = \left(tf_{sj}\right)_{l\times m} = TW \times FW^T = \begin{pmatrix} tf_{11} & tf_{12} & \cdots & tf_{1m} \\ tf_{21} & tf_{22} & \cdots & tf_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ tf_{l1} & tf_{l2} & \cdots & tf_{lm} \end{pmatrix} \tag{2}$$

Based on the work done, we store the subscripts of the failure modes identified in all titles in an array, then generate a collection of failure modes corresponding to all the problem titles. Again, the occurrence frequency of each failure mode is calculated.

According to the component subscript and failure modes subscript corresponding to each problem title, the component is associated with the failure mode, and the number of failure modes of each component is calculated accordingly. Furthermore, the CF matrix is established.

The CFMM algorithm is shown in Table 1. The application of this algorithm will be explained in the next section with an example.

As shown in the algorithm, the purpose of steps 1–11 is to determine which components are included in each title. In steps 12 and 13, the number of occurrences of each component is calculated as a statistical result. Similarly, the purpose of steps 14 to 23 is to identify the failure mode contained in each title. In steps 24 and 25, the number of occurrences of each failure mode is calculated as a statistical result.

According to Steps 8, 9, 20, and 21, we obtain the component subscript and failure mode subscript corresponding to each problem title. Then, through step 27, the components are associated with the failure modes, and the number of failure modes of each component is calculated accordingly. Furthermore, the CF matrix is established.

Case study

Experimental data

In this section, we take problem data of a car seat system from Company A as an example to verify the proposed method. The seat is an important part of the automotive interior. In addition to providing smooth operation and comfortable driving for the passengers, it must also have the function of ensuring the safety of the passengers. At the same time, some seats also have the function of heating, automatic adjustment, and other requirements to meet the individual needs of customers. The failure modes of the car seat components may have various effects; some may affect the appearance, some may affect the function, and even more pose a safety hazard. Therefore, some measures can be taken from the design stage to avoid problems by identifying the failure modes of each component of the seat and performing FMEA. The seat system includes front seat assembly, rear seat assembly, seat belt system, and child restraint system for a total of more than 300 components. Due to space limitations, this paper presents some common components in Table 2.

We obtained 11,677 problem records from the year 2010 to 2016 from the quality management information system of Company A, of which 568 are related to the seat. According to the title of these data, we delete the items that only contain simple content such as "problems," "problem," "seat problem." Meanwhile, some titles written in German are also deleted. After this process, the number of useful seat problems is 495. Each record includes the problem number, vehicle model, main module, title, description, creation date and other information. Before data mining, we construct the corresponding stopword list and obtain the acronym table commonly used in Company A.

The problem titles are refined short texts. The extraction of the failure modes and the construction of the CF matrix are based on the problem titles. In company A, there is a standard for the input of the quality problem title, usually "problem finder _ model _ project stage _ problem concise description." For example, in the problem title "FDP_F35_SE_Noise from right rear seat backrest unlocking as driving on the bumpy road", "FDP" indicates that the problem was discovered by a road test, "F35" indicates that the problem occurred on the F35 model, and "SE" indicates that the problem is mass production problem. Despite the input criteria, the problem finder will occasionally describe the problem in the way he is used to, resulting in reduced data quality. According to the statistics, 65% of the problem input of the seat module meets the standard, and the remaining problem title is a concise description of the problem. However, this does not affect the subsequent analysis of this paper. In the process of extracting the failure mode, this paper filters the information of the problem finder, model and project stage and only uses the concise description.

Failure mode extraction result

This section uses the WordNet-based approach described in "Standard failure mode set construction" section to build a standardized failure mode set. Table 3 presents a synonym set of standard failure modes. As shown in the table, the synonymous failure modes are combined, and a total of 17 groups are obtained. The number in parentheses after each failure mode indicates the support count for the failure mode. For most groups, the failure mode with the highest support count is considered the standard failure mode, which is the most commonly used expression for inputting the problem from different quality departments. For some groups, although some failure modes have the highest support count, the experts judge the other failure modes in the group as the standard failure mode. For example, in group 12, "aroma," "smell," and "odor" indicate that the seat emits an unusual smell. Although "smell" has the highest support count, "odor" is a more professional expression, so it is adopted as the standard failure mode of this group. The

Table 1 Component and failure mode association mining algorithm

Algorithm 1 CF matrix mining (CFMM)

Input: CW , FW , TW
Output: CF .

for $s \leftarrow 1$ **to** l **do** ▹ for each problem title
  for $i \leftarrow 1$ **to** n **do** ▹ for each component
    $p[i-1] \leftarrow TW(s,:) \times CW^T(i,:)$ ▹ The number of identical words in component and title.
  end for
  find $i \leftarrow \alpha$ s.t. $p[\alpha-1] \leftarrow max\, p[n]$
  if the number of the maximum value in $p[n]$ is greater than one **do**
    find $\alpha \leftarrow \alpha'$ s.t. $\sum_{\tau=1}^{e} cw_{\alpha'\tau} \leftarrow min \sum_{\tau=1}^{e} cw_{\alpha\tau}$ **then** ▹ the component that has the fewest words
    $x[s-1] \leftarrow \alpha'$
  else $x[s-1] \leftarrow \alpha$
  end if
  return $CX \leftarrow \{c_{x[0]}, c_{x[1]}, \cdots, c_{x[l-1]}\}$
  for each $c_{x[k]} \in CX$ **do**
    $c_{x[k]} \leftarrow c_{x[k]}.count++$
  for $j \leftarrow 1$ **to** m **do** ▹ for each failure mode
    $q[j-1] \leftarrow TW(s,:) \times FW^T(j,:)$ ▹ the number of identical words in failure mode and title
  end for
  find $j \leftarrow \beta$ s.t. $q[\beta-1] \leftarrow max\, q[m]$
  if the number of the maximum value in $q[m]$ is greater than one **do**
    find $\beta \leftarrow \beta'$ s.t. $\sum_{\tau=1}^{e} fw_{\alpha'\tau} \leftarrow min \sum_{\tau=1}^{e} fw_{\alpha\tau}$ **then** ▹ the failure mode that has the fewest words
    $y[s-1] \leftarrow \beta'$
  else $y[s-1] \leftarrow \beta$
  end if
  return $FY \leftarrow \{f_{y[0]}, f_{y[1]}, \cdots, f_{y[l-1]}\}$
  for each $f_{y[k]} \in FY$ **do**
    $f_{y[k]} \leftarrow f_{y[k]}.count++$
end for
return $D \leftarrow \{cf_{x[0]y[0]}, cf_{x[1]y[1]}, \cdots, cf_{x[l-1]y[l-1]}\}$
for each $cf_{x[k]y[k]} \in D$ **do**
  $cf_{x[k]y[k]} \leftarrow cf_{x[k]y[k]}.count++$
end for
for each $cf_{ij} \in CF_{nm}$ **do**
  If $cf_{ij} \in D$
    $cf_{ij} \leftarrow cf_{ij}.count++$
  else $cf_{ij} \leftarrow 0$
  end if
end for
return CF_{nm}

support count of the standard failure mode will be the sum of support count of the synonymous failure modes in each group. The standard failure modes of seat module presented in Table 3 are not only applicable in company A but are also an important reference for all vehicle manufacturers and corresponding seat suppliers.

After the failure mode mining of 495 seat quality problem records, a total of 57 types of failure modes were recognized. Due to space constraint, this paper only displays in Fig. 2 the failure modes that occurred more than six times. As shown in Fig. 2, “noise”, “wavy”, “defect”, “gap”, and “function defect” are the top five failure modes. This chart is a kind of Pareto Chart, the statistics in this chart will be a guide for the managers of quality management. According to Fig. 2, the quality managers will identify the main failure modes that occurred on seat module and perhaps take some priority measures to solve these failure modes to improve key performance indicator (KPI) such as defects per 100 units (DPU). In this way, we

Table 2 A portion of the seat assembly

System	Component
Front seat assembly	Seat rail/seat armrest/seat backrest/headrest/covers/finishers/seat heating/first aid box
Rear seat assembly	Supports/covers/headrest/center armrest/ski bag/finishers/seat heating
Seat belt system	Seat belt/belt height adjuster/belt tensioner/belt buckle/end fittings
Child restraint system	Child seat impact table/child seat height adjustment/child seat footrest/ISOFIX

Table 3 The Standard failure mode set

No	Synonymous failure mode set with support count	Standard failure mode
1	Squeaking (6), knocking (0), creaking (2), sound (0), rattle (11), noise (109)	Noise (128)
2	Failure (3), malfunction (6), defect (55)	Defect (64)
3	Wrinkle (12), wavy (55)	Wavy (67)
4	Move (2), movement (2), loose (22)	Loose (26)
5	Incorrect (0), fault (2), abnormal (2), wrong (13)	Wrong (17)
6	Lose (0), omitted (2), disappear (0), missing (8)	Missing (10)
7	Thermal (1), heating (7)	Heating (8)
8	Broken (3), damage(7)	Damage (10)
9	Thread (2), stitch (2), sewing (2), seam (6)	Seam (12)
10	Friction (0), rubbing (4), detrition (0)	Rubbing (4)
11	Shake (0), vibration (3)	Vibration (3)
12	Aroma (0), smell (4), odor (2), scent (0)	Odor (6)
13	Warning (0), alarm (2)	Alarm (2)
14	Pollutant (1), contamination (2)	Contamination (3)
15	Delamination (0), lamination (2)	Lamination (2)
16	Not parallel (1), parallelism (0), misalignment (2), tapered (0), wedge (0)	Not parallel (3)
17	Deformed (0), distortion (0), twist (2)	Deformed (2)

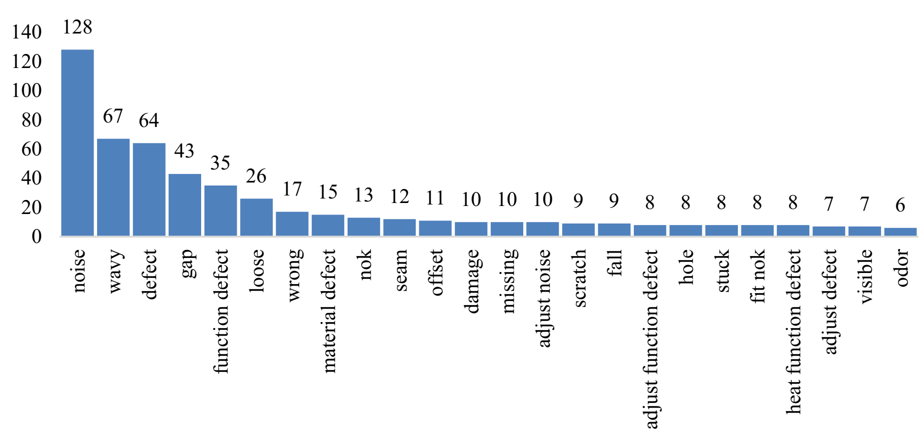

Fig. 2 The standard failure modes and their frequency

can organize and utilize resources such as personnel and equipment more effectively in quality management.

In addition, to compare the method of this paper with other methods, we adopt the clustering method and the FP-growth algorithm in RapidMiner Studio to extract the failure modes from the text. RapidMiner is a drag-and-drop graphical tool for machine learning, data mining, text mining, predictive analysis, and business analysis. It is a kind of tool embedded with algorithms such as K-Means which is a classical clustering algorithm and FP-growth which is a classical algorithm to extract frequent itemsets. Some other text mining tools or programming languages can also be used to deal with these tasks.

After the pretreatment process, we exploit the K-means algorithm for clustering and the squared Euclidean distance as a measure of distance between samples, which is the sum of quadratic differences overall attributes.

K-means clustering requires a prior determination of the cluster values K_{max}. There is no clear theoretical guidance on how to determine the K_{max}. Most scholars use the empirical rule for $K_{max} \leq \sqrt{n}$, where n is the number of data objects (Rezaee et al. 1998; Limwattanapibool and Arch-Int 2017). Therefore, according to the method and considering a total of 495 problem records, the predetermined category does not exceed 22 categories, which means $K = 22$.

Table 4 shows the clustering results for all quality problem text records. The descriptive terminologies in the table are representative terms for a cluster selected by RapidMiner according to the order of TF-IDF of these terms; the absolute count is the number of files in the cluster; the coverage is the number of documents in the cluster divided by the total number of documents in the collection. After judging, some of the different clusters are a type of failure mode, but they are divided into different clusters. For example, cluster 2, cluster 8, cluster 17, and cluster 18, are noise-type failure modes. Therefore, we combine these four clusters into one cluster and define it as the "noise" failure mode type. Finally, we merge the 22 clusters into 18 failure modes and add the same cluster counts as the count of the failure modes.

Furthermore, this paper also uses the FP-growth algorithm in RapidMiner to extract the frequent failure mode set. In this part, the FP-growth algorithm is used to extract frequent itemsets. To eliminate the infrequent itemsets, the minimum support was set to 0.02, where minimum support = (number of occurrences of an itemset)/(size of the example set). After manual pruning, RapidMiner extracted a total of 495 failure modes for 71 categories.

Table 4 Failure mode extraction results with the K-means clustering method

Cluster	Absolute count	Coverage (%)	Descriptive terminologies	Failure mode type	Failure mode count	New cluster number
Cluster 0	8	1.6	Malfunction, without, omit	Malfunction	8	C_0
Cluster 1	20	4.0	Loose, screw, easy, trim	Loose	20	C_1
Cluster 2	56	11.3	Noise	Noise	160	C_2
Cluster 8	90	18.2	Noise, squeak,, drive			
Cluster 17	11	2.2	Rattle, noise, biw			
Cluster 18	3	0.6	Guide, damage, noise			
Cluster 3	24	4.8	Gap, misalign, taper	Gap	39	C_3
Cluster 10	15	3.0	And, between, gap, seal			
Cluster 4	14	2.8	Not, accept, smell	Smell	14	C_4
Cluster 5	13	2.6	Wrong, decor, direct	Wrong	13	C_5
Cluster 6	43	8.7	Scratch, damage, stuck, touch	Scratch	43	C_6
Cluster 7	12	2.4	Cannot, open, adjust	Cannot open	12	C_7
Cluster 9	8	1.6	Hole, outer, close	Hole	8	C_8
Cluster 11	51	10.3	Wavy	Wavy	51	C_9
Cluster 12	15	3.0	Material, defect	Material defect	15	C_10
Cluster 13	22	4.4	Adjust, function, noise, defect	Adjust defect	22	C_11
Cluster 14	12	2.4	Wrinkle, area, edge	Wrinkle	12	C_12
Cluster 15	13	2.6	Fall, off, mechanism	Fall off	13	C_13
Cluster 16	10	2.0	Miss, label, inform	Missing	10	C_14
Cluster 19	12	2.4	Nok, fit, corner	Fit nok	12	C_15
Cluster 20	11	2.2	Offset, between, and, sit	Offset	11	C_16
Cluster 21	32	6.5	Function, defect, heat	Function defect	32	C_17

Table 5 Comparison of failure modes extracted by three methods

	FP-growth	K-means clustering	Our method
Failure mode categories	71	18	57

Table 5 shows the categories and number of failure modes extracted by these three methods. As we can see, our method extracts 57 failure modes categories, which can yield more failure mode categories than the clustering method but fewer than the FP-growth method. For the frequent itemsets extracted by FP-growth, there are some failure modes described with different words but the same failure mode. For example, FP-growth extracts "noise" with a support count of 109 and "rattle" with a support count of 11, which should be the same failure mode according to the domain experts. Another example will be "wavy" with support count of 55 and "wrinkle" with the support count of 12, but both of them means the wavy of the seat surface. In these 71 failure modes categories extracted by FP-growth, there are some synonyms, but they are not combined and standardized. Based on these failure mode set, the users will be confused when the select words to describe the failure modes. However, our method not only extracts frequent itemsets of failure modes but also builds up the synonyms of failure modes based on WordNet. At the same time, the problem titles are always written by a short text, the matrix represented by the vector space model will be sparse. Thus it is not easy to cluster the similar text in the same group by cluster algorithms, and sometimes different text will be clustered into one group. As shown in Table 5, "scratch," "damage," "stuck," and "touch" are clustered in one group by the clustering algorithm, which should be the different failure modes. At the same time, "noise" distributed in cluster 2, 8, 17, and 18. The clustering result is not so satisfactory. On the whole, our method can obtain a variety of high-quality failure mode sets.

The failure modes with support count greater than ten are selected and presented in Fig. 3 due to the limited space. The number of failure modes extracted by these three methods is compared. As shown in Fig. 3, most of the failure modes obtained by our method are the highest. However, for partial failure modes such as "noise," the clustering method obtains the largest number, mainly because the clustering result is not particularly accurate.

Component-failure mode matrix construction result

In this section, we use the CFMM algorithm in "Component-failure mode matrix mining algorithm" section to mine the association matrix between the seat components and the failure modes from the set of quality problem

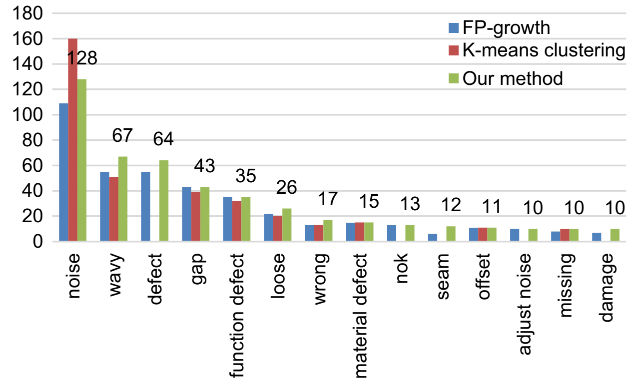

Fig. 3 Comparison of the number of failure modes obtained by the three methods

titles. The algorithm effectively identifies 110 component categories for a total of 495 seat components from all the quality problems. The identified useful failure modes are 57 categories. Similarly, the seat components that appear more than twice are presented in Fig. 4.

As shown in Fig. 4, "seat backrest," "seat," "seat belt," "seat headrest," and "seat cover" are the top-five most popular seat components. In the actual situation, the component described by "seat" is said to be the complete seat in the module structure. However, it is possible that some employees did not specify specific components when entering data.

As mentioned earlier, the number of components categories extracted is 110, and the failure mode category is 57. That is, the CF matrix is a 110*57-dimensional matrix. In this matrix, there are 266 nonzero items (CF combinations), and the total number of all failure modes that have occurred for all components is 495. Because the matrix obtained by the CFMM algorithm cannot be presented, we use Table 6 as an example to specify the CF matrix.

From the table, we can see that the matrix is a three-by-four dimensional matrix. In this matrix, there are five nonzero items, which means that the CF combinations have occurred five times. They are "seat belt broken," "seat belt noise," "seat belt stuck," "seat backrest wavy" and "cup holder noise." As shown in the table, the total number of failure modes for all components is eleven.

This paper also uses the FP-growth algorithm in RapidMiner to extract components and failure mode frequent itemsets. The minimum support is set to 0.02 to eliminate itemsets that do not occur frequently. After manual pruning, the FP-growth algorithm in RapidMiner extracts 50 categories of CF combinations. The sum of the support count of all combinations is 188. Among these 50 categories of CF combinations, there are 14 components categories and 27 failure modes categories. According to this result, the CF matrix whose size will be 14*27 can be constructed. Table 7

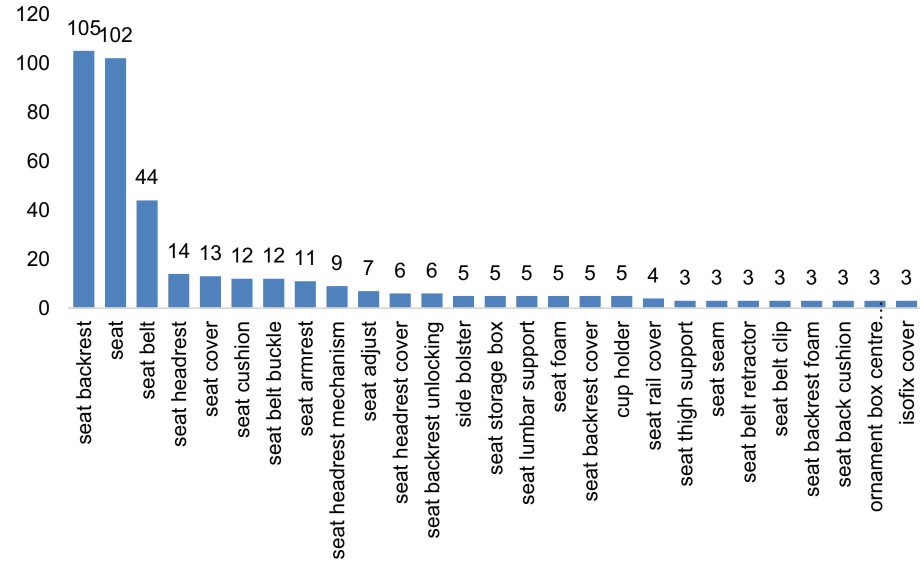

Fig. 4 Components identified by the CFMM algorithm

Table 6 An example of a CF matrix

Components: 3 Failure modes: 4 Problem sources: 11	Wavy	Broken	Noise	Stuck
Seat belt	0	1	2	1
Seat backrest	3	0	0	0
Cup holder	0	0	4	0

Table 7 Comparison of CF matrix identified by CFMM and FP-growth algorithm

	CF matrix dimension	CF combination categories	CF combination quantity
CFMM	110×57	266	495
FP-growth	14×27	50	188

presents the differences between the CF combination mined by the CFMM method and the FP-growth algorithm.

Furthermore, we compare the number of each combination of CF extracted by the CFMM algorithm and FP-growth algorithm. Because the CF matrix obtained by CFMM is too large, this paper cannot present all CF combinations. Therefore, the CF combinations that occur more than three times are presented in Fig. 5.

As shown in Fig. 5, the number of each CF combination obtained by the CFMM algorithm is higher than the corresponding number obtained by the FP-growth algorithm, which shows that the CFMM algorithm proposed can extract the CF matrix more effectively.

To visualize the relationship between the seat components and the failure modes, we use the data analysis platform Gephi to present the CF matrix in the form of a network diagram. To better present the effect, we delete the isolated components or failure modes in the matrix that are not associated with other points. As shown in Fig. 6, the components and the failure mode nodes are mixed to form a complex connection network. The larger the node and the number, the more times the component or failure mode occurs. In this network diagram, two directly connected nodes must be a component and a failure mode. Two failure modes or two components cannot be connected by one edge. The edge between two nodes indicates the number of times the component experienced the failure mode. The thicker the edge, the more times the component has experienced the failure mode.

Through the above network diagram, quality management personnel can have a more macroscopic and direct understanding of the failure modes that appear in the components. It is very important for supplementing the DFMEA at the enterprise level. In addition, the quality management personnel can further drill down according to Fig. 6 to analyze the composition of the failure modes of each component. Therefore, allocating more quality management-related resources to those key components and failure modes. According to the network diagram, this paper takes the seat backrest with the most occurrence as an example for further analysis. The drill-down analysis result is shown in Fig. 7. The three most common failure modes on the seat backrest are "wavy," "noise" and "gap." For this case, the enlightenment is that quality management personnel should focus their attention on these three categories of failure modes on the seat backrest.

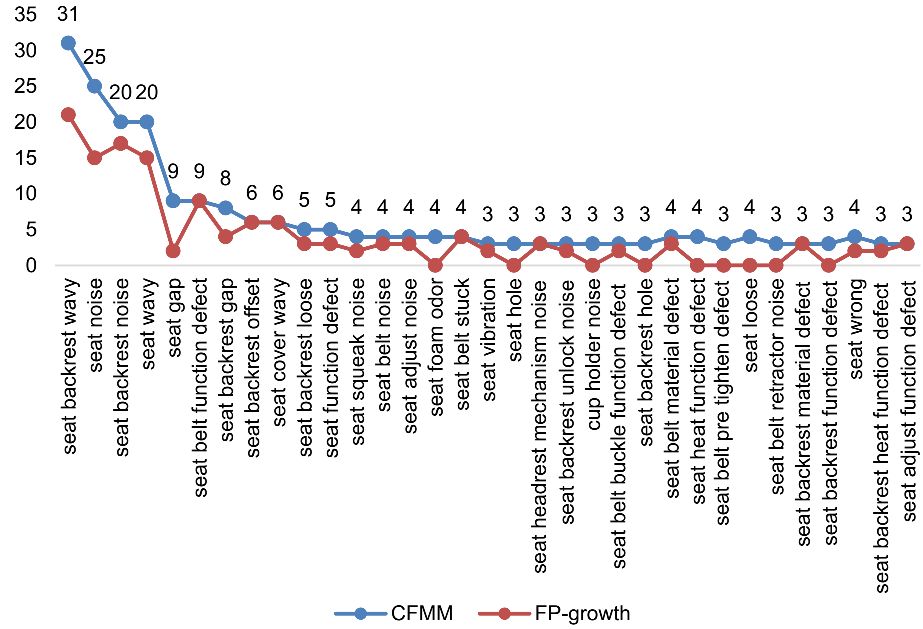

Fig. 5 CF combination extraction results comparison

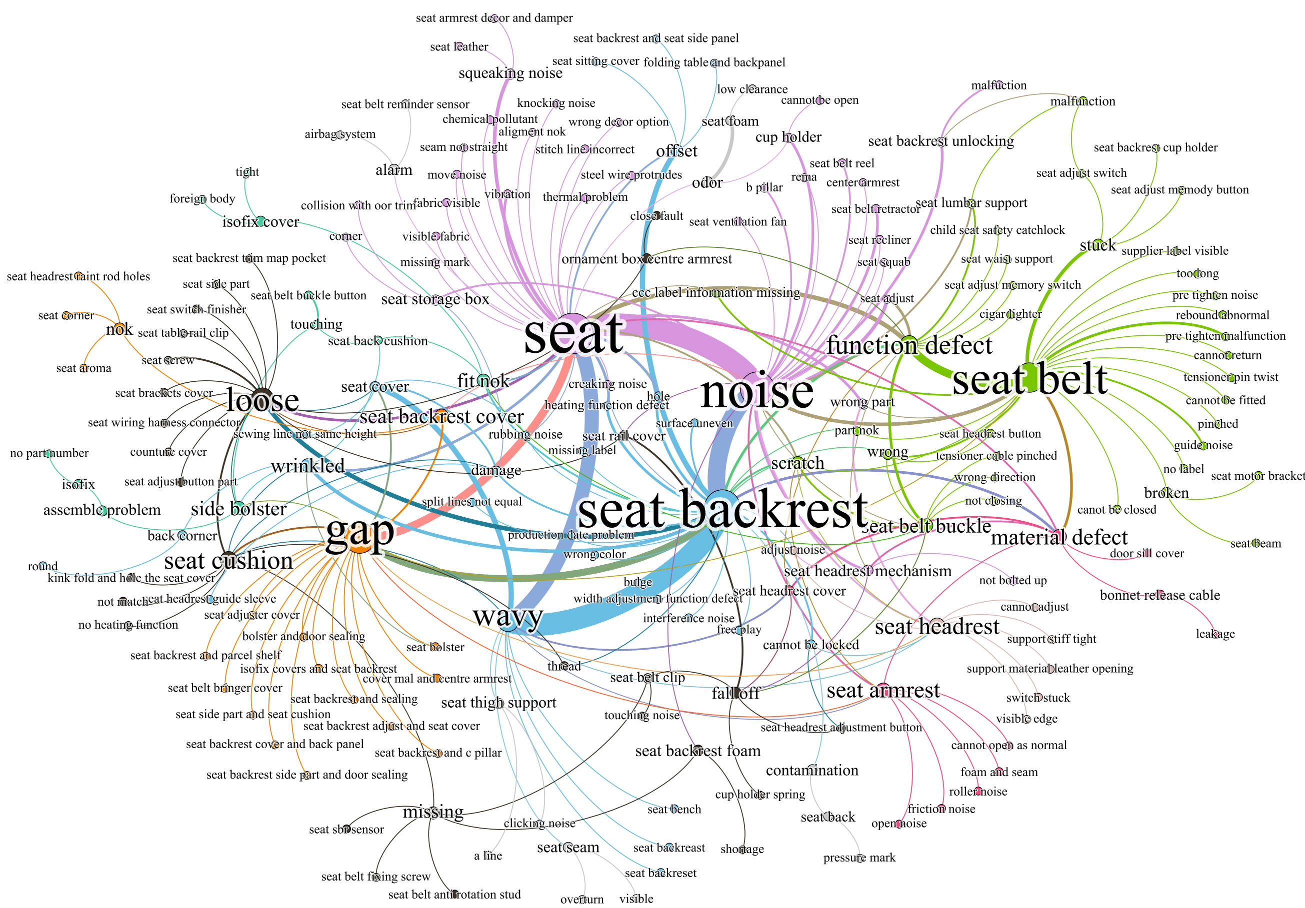

Fig. 6 Network map between seat components and failure modes

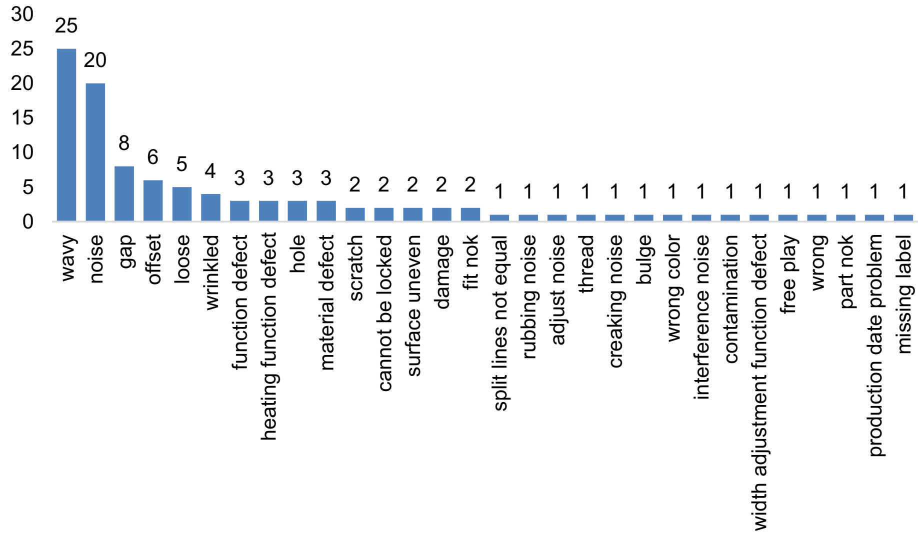

Fig. 7 Seat backrest failure mode drill-down analysis

Conclusions and future work

In this study, we developed a novel systematic approach of effectively establishing the design failure-mode and effects analysis upon the standardization of failure modes and CF matrix mining, to overcome some significant shortcomings of existing methods. Different departments use different vocabulary when describing the same failure mode, so building standard failure mode vocabulary can improve communication between departments and improve people's understanding of failure modes. However, manually building standard failure mode vocabulary is a time-consuming and labor-intensive process. At the same time, the CF matrix is an essential source of knowledge for DFMEA, while manually extracting CF matrix from large amounts of different documents is also not an easy task.

The focused DFMEA here is a fundamental tool for improving quality and enhancing the reliability of products. DFMEA is an effective weapon commonly used in product design and development to take full account of the problems involved in the production, delivery, and use of products, to bring all possible problems into the scope of prevention, and to prepare preventive means in advance. The creation of DFEMA first needs to know which failure modes have occurred in the product components. Component-failure mode matrix is an important source of knowledge in this process. On the one hand, the designers who create DFMEA are far away from the production process, lack of understanding of the product quality problems that may occur in the production process, and the data of product quality problems scattered in the production process form an information isolated island, which is difficult for designers to use. On the other hand, the employees who record product quality problems often adopt according to their own habits when describing the same problem. Different words are used to describe the failure mode, which results in the ambiguity of the designer's perception of failure mode. To solve these problems, this paper proposes a method to mine failure modes from quality data recorded in a large number of production processes and construct a standard failure mode library as a general language for problem description between different departments. On this basis, we use a large number of product quality data to construct component-failure mode matrix automatically and use it as a knowledge reference for designers to create DFMEA.

In response to these problems, this paper first extracted a list of frequent failure modes from problem-solving data by Apriori algorithm, then based on WordNet, find the synonymous failure modes from the list, and then build the standard failure mode vocabulary. Based upon the standard failure mode vocabulary and the existing component set, the quality problem title with implied failure mode information was used as a link, and the CFMM algorithm was examined to construct the CF matrix automatically. This paper adopted a car company's seat module as an example to analyze the results of the standard failure modes and the effect of the CFMM algorithm. The results showed that the failure mode extraction method with standardized features could extract the failure mode better than the FP-growth and K-means clustering methods. At the same time, the CFMM algorithm could extract more CF combinations and build a richer set of CF matrices than the FP-growth method. Although each industry has different domain characteristics, the method in this paper is applicable not only to the manufacturing industry but also to other fields that need to use FMEA to ensure product and system reliability.

Our theoretical contribution can be in large part reflected in the innovative component-failure mode matrix (CFMM) algorithm used in the DFMEA construction process. The proposed method has several advantages over the existing methods: (1) In the construction of standard failure modes, the FP-growth approach does not standardize failure modes, and many essentially same FMs are computed into different FMs, resulting in more significant errors. While the K-means clustering method uses rough fields in FM recognition and extraction, and many of the FMs that should have been included are eventually omitted with poor accuracy. (2) For developing CFMM, the FP-growth algorithm only seeks the correlation between FM and components with significant frequency in frequent itemsets, and the coverage of FMs is relatively narrow, whereas the CFMM algorithm in this paper covers both significant and insignificant FM in frequent itemsets, and constructs the correlation matrix between standard failure modes and components more completely with high accuracy.

This paper mined the quality problem title recorded in text form, which serves as a concise information representation that provides important information about failure modes. However, in the text of other records such as quality problem descriptions, partial failure mode information is also included, and it is possible that one description contains multiple failure modes. Identifying failure modes from longer texts and building a CF matrix will continue to be studied in the future. Due to space limitations, this paper only studied the relationship between components and failure modes. However, in the text of quality problems, the causal relationship between failure mode and cause may also be implied, and this will also be an important source of knowledge for FMEA. Therefore, the relationship between failure mode and cause may also be one of the problems studied in future research.

Acknowledgements This research has received funding from National Natural Science Foundation of China (NSFC) (Grant No. 71871041). The authors would like to thank the Institute of systems engineering of the Dalian University of Technology and BMW Brilliance Automotive Ltd. for supporting the research. Special thanks to Dr. Longfei He from Tianjin University for his important participation and substantive contribution in this research by providing valuable ideas and constructive suggestions and revising the whole paper.

References

Arunajadai, S. G., Stone, R. B., Tumer, I. Y., & Clancy, D. (2002). *A Framework for creating a function-based design tool for failure mode identification.* Paper presented at the ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Montreal, Canada.

Arunajadai, S. G., Uder, S. J., Stone, R. B., & Tumer, I. Y. (2004). Failure mode identification through clustering analysis. *Quality and Reliability Engineering International, 20*(5), 511–526. https://doi.org/10.1002/qre.663.

Asan, U., & Soyer, A. (2016). Failure mode and effects analysis under uncertainty: A literature review and tutorial. In C. Kahraman & S. Yanik (Eds.), *Intelligent decision making in quality management: Theory and applications* (pp. 265–326). Switzerland: Springer.

Baroni, M., & Bisi, S. Using cooccurrence statistics and the web to discover synonyms in a technical language. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004* (pp. 1725–1728): European Language Resources Association.

Beksinska, M., Joanis, C., Manning, J., Smit, J., Callahan, M., Deperthes, B., et al. (2007). Standardized definitions of failure modes for female condoms. *Contraception, 75*(4), 251–255. https://doi.org/10.1016/j.contraception.2006.10.003.

Blondel, V. D., & Senellart, P. P. Automatic extraction of synonyms in a dictionary. In *Proceedings of the Society for Industrial and Applied Mathematics Workshop on Text Mining, Arlington, Virginia, USA, 2002* (pp. 7–13).

Brook, Q. S. (2006). *Six Sigma and Minitab: A complete toolbox guide for all Six Sigma practitioners* (2nd ed.). London: QSB Consulting.

Chang, W., Kai, M. T., & Lim, C. P. (2015). Clustering and visualization of failure modes using an evolving tree. *Expert Systems with Applications, 42*(20), 7235–7244.

Chen, L., & Nayak, R. (2007). *A case study of failure mode analysis with text mining methods.* Paper presented at the International Workshop on Integrating Artificial Intelligence & Data Mining, Darlinghurst, Australia.

Cheng, T., Lauw, H. W., & Paparizos, S. (2012). Entity synonyms for structured web search. *IEEE Transactions on Knowledge and Data Engineering, 24*(10), 1862–1875.

Collins, J. A., Hagan, B. T., & Bratt, H. M. (1976). The failure-experience matrix: A useful design tool. *Journal of Engineering for Industry, 98*(3), 1074–1079.

Deore, R. Y. (2015). An ontology text mining to conversion of unstructured to structure text in D-Matrix. *Indian Journal of Scientific Research, 6*(1), 47–48.

Fellbaum, C. (2000). WordNet: An electronic lexical database. *Language, 76*(3), 706–708.

Goel, A., & Graves, R. J. (2007). *Using failure mode effect analysis to increase electronic systems reliability.* Paper presented at the 30th International Spring Seminar on Electronics Technology (ISSE), Cluj-Napoca, Romania.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Amsterdam: Elsevier.

Hearst, M. A. (1999). Untangling text data mining. *Proceedings of ACL-99*, 3-10.

ISO/IEC-15026-1 (2013). Systems and software engineering-Systems and software assurance-Part 1: Concepts and vocabulary. Switzerland: ISO/IEC JTC 1/SC 7 Software and systems engineering subcommittee.

James, A. T., Gandhi, O. P., & Deshmukh, S. G. (2017). Knowledge management of automobile system failures through development of failure knowledge ontology from maintenance experience. *Journal of Advances in Management Research, 14*(1), 1–21.

Jenifa, J., & Balachander, T. (2015). Survey on fault dependency matrix construction using ontology based text mining. *International Journal of Advanced and Innovative Research, 4*(3), 347–350.

Kai, M. T., Jong, C. H., & Lim, C. P. (2015). A clustering-based failure mode and effect analysis model and its application to the edible bird nest industry. *Neural Computing and Applications, 26*(3), 551–560.

Khilwani, N., & Harding, J. A. (2016). Managing corporate memory on the semantic web. *Journal of Intelligent Manufacturing, 27*(1), 101–118.

Kim, S., & Yoon, J. (2015). Link-topic model for biomedical abbreviation disambiguation. *Journal of Biomedical Informatics, 53,* 367–380.

Limwattanapibool, O., & Arch-Int, S. (2017). Determination of the appropriate parameters for K-means clustering using selection of

region clusters based on density DBSCAN (SRCD-DBSCAN). *Expert Systems, 34*(3), 1–11.
Liu, H., Chen, Y., You, J., & Li, H. (2016). Risk evaluation in failure mode and effects analysis using fuzzy digraph and matrix approach. *Journal of Intelligent Manufacturing, 27*(4), 805–816.
Liu, L., Fan, D., Wang, Z., Yang, D., Cui, J., Ma, X., et al. (2017). Enhanced GO methodology to support failure mode, effects and criticality analysis. *Journal of Intelligent Manufacturing*, *2017*, 1–18. https://link.springer.com/article/10.1007%2Fs10845-017-1336-0.
Mangnoesing, G., Bunningen, A. V., Hogenboom, A., Hogenboom, F., & Frasincar, F. (2012). *An Empirical Study for Determining Relevant Features for Sentiment Summarization of Online Conversational Documents.* Paper presented at the International Conference on Web Information Systems Engineering, Berlin.
Mendhe, S. P., & Hande, K. N. (2017). Graphical analysis on text mining unstructured data using D-matrix. *International Journal on Recent and Innovation Trends in Computing and Communication, 5*(6), 1412–1416.
Meng, L., Wang, H., Xue, Y., & Ma, L. (2017). Research on automatic generation of software failure modes. *Journal of Frontiers of Computer Science and Technology, 12*(12), 1–11.
Miao, J., Zhang, Q., & Zhao, J. (2008). Chinese automatic text categorization based on article title information. *Computer Engineering, 34*(20), 13–15.
Muller, P., Hathout, N., & Gaume, B. Synonym extraction using a semantic distance on a dictionary. In *Proceedings of the first workshop on graph based methods for natural language processing, 2006* (pp. 65–72). Association for Computational Linguistics.
Pfeifer, T. (2002). *Quality management: Strategies, methods, techniques*. München: Hanser Verlag.
Rajpathak, D. G., & De, S. (2016). A data- and ontology-driven text mining-based construction of reliability model to analyze and predict component failures. *Knowledge and Information Systems, 46*(1), 87–113.
Rajpathak, D. G., & Singh, S. (2014). An ontology-based text mining method to develop D-matrix from unstructured text. *IEEE Transactions on Systems Man & Cybernetics Systems, 44*(7), 966–977.
Rezaee, M. R., Lelieveldt, B. P. F., & Reiber, J. H. C. (1998). A new cluster validity index for the fuzzy c -mean. *Pattern Recognition Letters, 19*(3–4), 237–246.
Roberts, R. A., Tumer, I. Y., Stone, R. B., & Brown, A. F. (2003). *A Function-Based Exploration of JPL's Problem/Failure Reporting Database.* Paper presented at the ASME International Mechanical Engineering Congress and Exposition.
Schneider, H. (2003). Failure mode and effect analysis: FMEA from theory to execution. *Technometrics, 38*(1), 80.
Singh, S., Holland, S. W., & Bandyopadhyay, P. (2010). *Trends in the development of system-level fault Dependency matrices.* Paper presented at the Aerospace Conference, Big Sky, MT, USA.
Stamatis, D. H. (1995). *Failure mode and effect analysis: FMEA from theory to execution*. Milwaukee: ASQ Quality Press.
Thombare, T. R., & Dole, L. (2015). D-matrix: Fault diagnosis framework. *International Journal of Innovative Research in Computer and Communication Engineering, 3*(3), 1740–1745.
Tumer, I. Y., Stone, R. B., & Bell, D. G. Requirements for a failure mode taxonomy for use in conceptual design. In *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm, 2003*.
Tumer, I. Y., & Stone, R. B. (2003). Mapping function to failure mode during component development. *Research in Engineering Design, 14*(1), 25–33.
Wang, T., & Hirst, G. (2012). Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering, 18*(3), 313–342. https://doi.org/10.1017/S1351324911000210.
Wang, M., Wang, B., & Tang, X. (2011). Optimal component subset selection method based on component-failure knowledge. *Computer Integrated Manufacturing Systems, 17*(2), 267–272.
Wani, M. F., & Jan, M. (2006). *Failure Mode Analysis of Mechanical Systems at Conceptual Design Stage.* Paper presented at the ASME 8th Biennial Conference on Engineering Systems Design and Analysis, Torino, Italy.
Wijayasekara, D., Manic, M., & Mcqueen, M. Vulnerability identification and classification via text mining bug databases. In *IECON 2014—40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, 2014* (pp. 3612–3618). IEEE.
Wu, C., Chen, T., Jiang, R., Ning, L., & Jiang, Z. (2017). A novel approach to wavelet selection and tree kernel construction for diagnosis of rolling element bearing fault. *Journal of Intelligent Manufacturing, 28*(8), 1847–1858.
Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., Song, M., et al. (2015). A preliminary study of clinical abbreviation disambiguation in real time. *Applied Clinical Informatics, 6*(2), 364–374. https://doi.org/10.4338/Aci-2014-10-Ra-0088.
Xu, Z., Dang, Y., & Munro, P. (2018). Knowledge-driven intelligent quality problem solving system in the automotive industry. *Advanced Engineering Informatics, 38,* 441–457.
Yates, A., & Etzioni, O. (2009). Unsupervised methods for determining object and relation synonyms on the web. *Journal of Artificial Intelligence Research, 34,* 255–296.