Expert Systems With Applications 92 (2018) 206–215

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Knowledge extraction and visualization of digital design process

Jiwon Yang[a], Eunji Kim[a], Minhoe Hur[a], Sungzoon Cho[a,*], Myungbin Han[b], Iksang Seo[b]

[a] *Department of Industrial Engineering, Seoul National University, Seoul 151–742, Korea*
[b] *Engineering Quality Verification Team, Hyundai Motor Company, 150 Hyundaiyeonguso-ro, Namyang-eup, Hwaseong, Gyeonggi-do 445–706, Korea*

ARTICLE INFO

Article history:
Received 4 July 2016
Revised 2 September 2017
Accepted 4 September 2017
Available online 14 September 2017

Keywords:
Knowledge extraction
Visualization
Digital design

ABSTRACT

After digitally designing components of vehicles, a design team creates a virtual manufacturing environment that resembles actual manufacturing facilities. During this digital pre-assembly process, a review team examines each component, and records its problems and requirements in part verification reports. Once these reports are delivered to specific design team responsible for each part, the design team can make appropriate adjustments to their designs. This digital pre-assembly process can evaluate and prevent flaws in design prior to actual manufacturing, improving production quality and reducing manufacturing cost. As these reports are written in free text form, they, however, are not fully utilized for understanding problems arising from the design process. This paper proposes a method of applying text mining techniques on verification reports to extract insights for quality improvement. In this paper, following three text mining approaches are proposed: (1) Extracting n-grams for text preprocessing and constructing domain ontology; (2) Extracting meaningful insights from text preprocessing; (3) Creating intuitive visual tools to understand the extracted insights. The proposed method is applied on approximately 140,000 reports, and is validated through the quality of the answers obtained for the questions posed by the domain experts. The proposed method successfully extracts useful information from the text database, and provides intuitive graphical interface, thereby satisfying the need of the domain experts. This paper proposes a systematic framework of transforming huge amount of raw text data into intuitive visualization. Through this framework, meaningful knowledge can be extracted, analyzed and shared to improve the quality of the products. Main contribution of our paper is that it proposes a framework for knowledge extraction from pre-assembly process. Not only does it systematically arrange the data, but it also combines various data sources and creates a knowledge system to improve efficiency of the design process.

1. Introduction

Digital pre-assembly is a process of virtually assembling an item using the designs of each component (Baba & Nobeoka, 1998). Through applying text mining techniques to verification reports, which contain the details of the problems presented during the digital pre-assembly process, we aim to extract meaningful knowledge for quality improvement. Through virtually assembling the components of vehicles based on CAD design data, digital pre-assembly is intended to detect and prevent design flaws prior to actual manufacturing. Fig. 1 describes this process in detail. Once a review team detects problems from the design data, it records the problem descriptions and requirements in the verification reports as texts. The report contains both design related information (part types, responsible design team and requirements) and evaluation related information (responsible review team, problem descriptions and requirements). As represented by (2) in Fig. 1, the verification reports are subsequently delivered to corresponding design team responsible for developing each part. Based on the reports, the design team revises its design, and appends the details of its modification to the reports. As shown by (3) in Fig. 1, the revised reports are transferred back to the review team. After re-evaluating the revised design, the review team will either require additional modifications or complete the verification process by saving the revised reports in the database. In Fig. 1, review team I101 performs evaluation on Type_99 car, and delivers the reports to corresponding design team D135 and D122, respectively.

For efficient design process and defect prevention prior to manufacturing, each report generated during the digital pre-assembly process is stored in a database. Table 1 shows few examples of verification reports. Written as free text, each report contains

* Corresponding author.
E-mail addresses: ad1392@snu.ac.kr (J. Yang), eunjikim@dm.snu.ac.kr (E. Kim), dninb.kr@gmail.com (M. Hur), zoon@snu.ac.kr (S. Cho), hmbin@hyundai.com (M. Han), kahn5@hyundai.com (I. Seo).

http://dx.doi.org/10.1016/j.eswa.2017.09.002
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

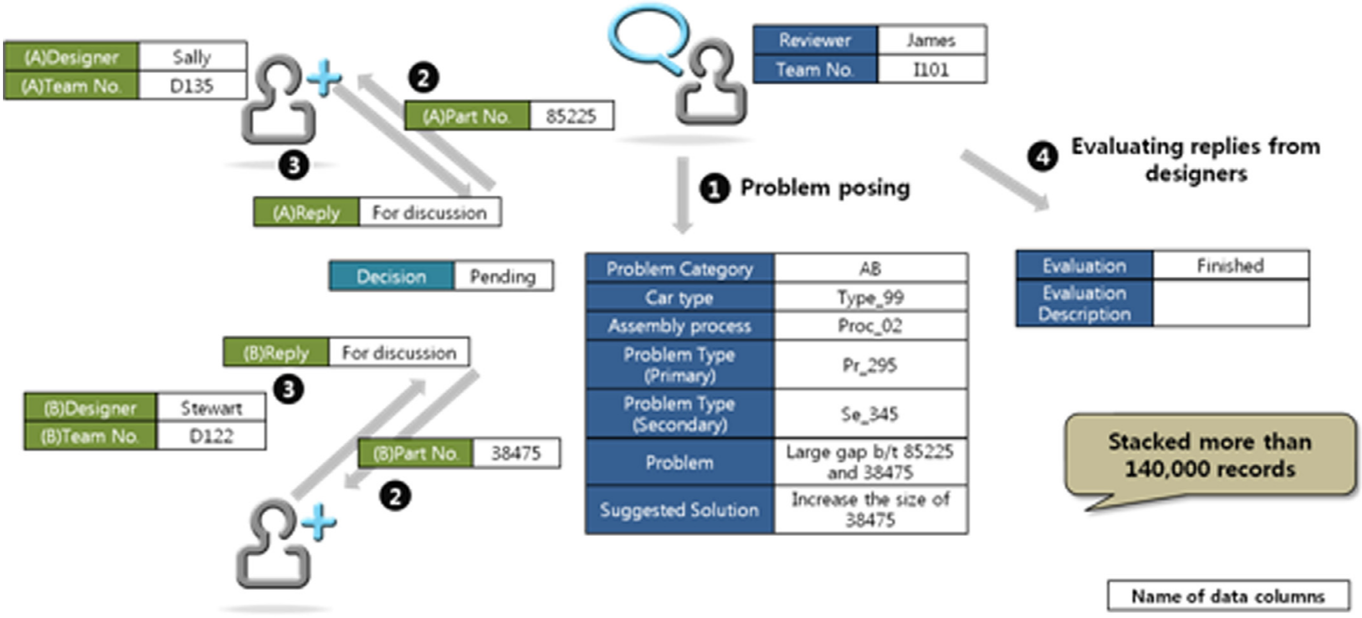


Fig. 1. Digital pre-assembly verification process.

Table 1
Problem/solution entries in the pre-assembly verification reports.

Problem	Suggested solution
Occurrence of penetration	Modify the design
Penetration issue	Avoid clash (Change the trim line)
After thickness(Amm) Less gap	Maintain a min gap of Bmm
Edge to fillet gap is less	Need MIN Cmm gap
Fillet to flange gap is less	Maintain MIN Dmm gap
...	...

the details of the problems posed by the review team and the solutions suggested by the design team.

Through systematically analyzing the collected verification reports, we can understand which parts cause most frequent problems, which review team poses these problems, and whether these problems are resolved or not. These insights can be applied in improving the quality of the manufactured parts. However, it is challenging to extract these insights from the reports as they are written in free text form (Soderland, 1999). In order to investigate the most frequent problem caused by specific parts, we have to search for appropriate reports using carefully selected search words, and manually read all of these reports from the system. Therefore, the current approach requires significant amount of time and manual labor, restricting the analysis to few limited number of reports. The verification reports analyzed in this paper have been collected over 6 years, containing approximately 140,000 reports for 100 different types of vehicles, design teams and review teams. Therefore, manually extracting insights for quality improvement from such huge and diverse database poses a great challenge. Furthermore, difference in terminologies used by various review teams and design teams presents an additional obstacle. For example, a term “shock absorber,” depending on the review team, are written as ‘SHK ABS’, “S/ABS” or “absorber.” These variations for identical terminologies not only prohibit information from being shared amongst different teams, but also cause great difficulties in extracting useful knowledge and insights.

This paper proposes a method of applying text mining techniques to extract meaningful information and knowledge from the verification reports, and presents them through intuitive visualization. Knowledge extraction from the text data consists of two steps. First, domain specific ontology, containing words related to the parts of the vehicles, must be created (Kietz, Volz, & Maedche, 2000). Providing deeper understanding than semantic reasoning, it must contain a knowledge structure that captures relationships between words or concepts. In order to extract concepts from the data, we tokenize our text data. Applying association rule mining technique, n-gram extraction is used for finding collocations, a series of words that represents a single concept. We then cluster similar concepts to represent them as a single concept, and label this relationship. These extracted concepts are subsequently linked to higher-level concepts, consisting of words such as “problem”, “cause”, “requirements”, “part” and “location.”

Second step of knowledge extraction involves text preprocessing with generated domain ontology. Through filtering step of preprocessing, unnecessary characters and signs are removed from the verification reports (Yao & Ze-wen, 2011). Once filtering step is completed, stemming process follows (Lovins, 1968). Although stemming is usually accomplished by the part-of-speech (POS) stemming, this conventional approach is problematic in our case as our free text reports contain various domain specific terms, spelling errors and ambiguous grammar structures. Consequently, constructed domain ontology is applied for stemming. Using the ontology, we replace the words with their corresponding concepts stored in it. If these concepts are affiliated with higher-level concepts, these higher-level concepts will replace the original words.

Finally, preprocessed data will be transferred to a visualization tool. As a fundamental step for understanding given data, data visualization arranges the data such that a user can visually recognize insightful patterns from the data. It includes both simple techniques for directly conveying information (bar chart, plot chart and line chart) and complex methods for in-depth understanding (drill-down and tree-map). Our research fully utilizes various aspects of data visualization.

This paper proposes a systematic framework of transforming huge amount of raw text data into intuitive visualization. Through this framework, meaningful knowledge can be extracted, analyzed and shared to improve the quality of the products. For a design team, the proposed framework can eliminate the need of manually searching through a database, saving both time and cost.

Furthermore, their design quality will improve as they now have easy access to information from the past. For a review team, they can examine how their reviews have been processed in the past, thereby creating more effective review delivery system.

The rest of this paper is structured as follows. In Section 2, we discuss various applications of text mining in manufacturing industry and their implications. Section 3 proposes our framework of creating domain ontology and ontology-driven text preprocessing. In Section 4, we will visualize the results acquired from our framework. We conclude in Section 5 with a summary of our research and directions for future works.

2. Related work

Ranging from process optimization, quality improvement to fault prediction (Choudhary, Harding, & Tiwari, 2009), data mining has been applied in various aspects of manufacturing industry. Through this technique, useful insights can be extracted, facilitating objective decision-making process. Previous applications, however, analyze structured data generated from sensors. Recently, unstructured data such as text has also been widely used for extracting insights in the manufacturing industry.

Automobile industry, as one of major manufacturing industries, has also been applying data mining techniques to analyze its data. Application of data mining in automobile industry often deals with detecting, predicting or analyzing malfunctions within vehicles. Recently, this narrow focus has expanded towards analyzing various stages in automobile production. Depending on the timing of data collection, automobile production can be divided into design phase, assembly phase and post-production phase. In post-production phase, data mining is applied to resolve issues caused while customers are driving finished vehicles. For example, Chougule, Rajpathak, and Bandyopadhyay (2011) applies data mining to customers' claim data in order to analyze and resolve problems with the customers' vehicles. During assembly phase, text data generated during the process has been analyzed to solve issues related to assembling various components of vehicles (Chakrabarty, Chougule, & Lesperance, 2009). As a most fundamental step, text analysis during design phase aims to predict potential problems in subsequent phases. As data analysis during this phase focuses on the earlier part of the entire manufacturing process, it can provide most significant quality improvement with relatively less effort and cost.

Furthermore, text mining is establishing its importance in manufacturing industries as explored by numerous papers. Lee et al. (2014) apply text mining to extract knowledge from inspection reports of marine structures. Through discovered insights, they systematically analyze and visualize complex relationships between various problem types. As huge amount of inspection reports are generated while constructing marine structures, it is impossible to analyze different types of issues and their relationships manually. Inspection reports, as input data, contain short free-text about inspection results and solutions to discovered problems. Their proposed method is composed of text preprocessing, document clustering and visualization. As the reports contain various technical terminologies, text preprocessing is essential in converting the free-text data into structured text data. With self-organized map (SOM), the reports are clustered and visualized, through which each cluster's keyword is used to represent a problem type. By applying principal component analysis (PCA) on concept linkage graph to produce two dimensional representation, relationships between the concepts (problem types) are visualized. Through this process, they discover most frequently occuring problem types and their unique characteristics for each marine structure. Not only do the new insights improve the existing manufacturing process, they also greatly reduce the time and cost required for filling the inspection reports

In semiconductor industry, text mining techniques are applied to analyze the relationship between event logs and yield (Kim, 2012). Although semiconductor manufacturing process has been automated, low yield of unknown cause is not automatically detected. When such phenomenon occurs, engineers have to classify the event logs produced during the manufacturing process and analyze the impact of each event on the yield. As such manual analysis requires significant amount of time and the results can be easily influenced by the expertise of the engineers, this paper propose a model that automatically classify and analyze text data. The overall model uses semiconductor yield data and event log data as input data. To find groups of meaningful keywords from the event logs, the paper utilizes both the engineers' experience and TF-IDF. Through combining these keywords from the log data with the yield data, this paper creates and evaluates various models such as linear regression, stepwise linear regression and neural network. Through these models, the paper confirms that fault detection related keywords have significant influence in the models, and successfully creates an automatic system that can predict low yield from semiconductor manufacturing process.

In automobile industry, ontology-guided knowledge retrieval system in an assembly environment has previously been proposed (Chakrabarty et al., 2009). When problems occur during the automobile assembly, engineers resort to keyword-based search through a database to find adequate solutions. During keyword search, correct results are retrieved only when the search keyword matches exactly with the words or phrases within the database. As finding matching keyword is extremely challenging, this paper proposes a search system that retrieves desired results without exactly matching keywords. As input data, it uses Variation Reduction Advisor (VRA) system, which is a database of problems encountered during the assembly process and their solutions. Creating ontology often relies heavily on manual efforts, prone to human errors and requiring considerable amount of time and cost. However, this paper proposes a system that automatically creates synonyms in order to reduce this manual effort. Furthermore, its system retrieves the result through connecting similar phrases instead of using conventional word-based ontology. Its proposed ontology-based search can be expanded for identifying problems and their solutions from various information retrieval systems.

Besides these applications, text mining has been applied in various fields of automobile industry such as fault diagnosis (Rajpathak, Siva Subramania, & Bandyopadhyay, 2012; Rajpathak & Singh, 2014), data acquisition from fault diagnosis (Huang, McMurran, Dhadyalla, & Jones, 2008) and service and repair system (Chougule et al., 2011; Khare & Chougule, 2012). However, there are no previous studies, in which text mining is directly applied to pre-assembly process (Table 2).

A more rigorous investigation on the existing methods, such as comparison of previous approaches in terms of pros and cons, is given in Table 3.

During recent five years, many papers have actively utilized text data collected within a company to extract meaningful knowledge. However, the scope of these papers usually focuses on preprocessing or analyzing text data instead of utilizing the results or creating a system for the entire process.

Main contribution of our paper is that it proposes a framework for knowledge extraction from pre-assembly process. Not only does it systematically arrange the data, but it also combines various data sources and creates a knowledge system to improve efficiency of the design process.

Table 2
Applications of text mining in manufacturing industry.

Research	Category	Goal	Algorithms
Our work	Pre-assembly process	Knowledge discovery from the design verification reports for quality control	Ontology, Association Rule Mining
Rajpathak and Singh (2014)	Fault diagnosis	Fining relationship between observable symptoms and failure modes	Ontology, Association Rule Mining
Rajpathak et al. (2012)	Data acquisition from fault diagnosis	Solve the data quality issues: formalizing free texts	Ontology
Huang et al. (2008)	Fault diagnosis	Guiding off-line vehicle fault diagnosis with probabilities of root causes	Bayesian Belief Network
Chougule et al. (2011)	Service and repair in an automotive domain	Identifying causes and solutions of customer dissatisfaction in repair shop	Association Rule Mining, Case Based Reasoning
Khare and Chougule (2012)	Service and repair in an automotive domain	Finding the proper repairs on the symptoms from the service documents	Association Rule Mining, Anomaly Detection
Chakrabarty et al. (2009)	Assembly process	Retrieving relevant solutions for the problems in the assembly process	Ontology, Information Retrieval
Lee et al. (2014)	Assembly process	Knowledge discovery in inspection reports of marine structure	Self-Organized Map
Kim (2012)	Assembly process	Analysis on engineers process event log in semiconductor manufacturing	Stepwise Linear Regression

Table 3
Comparison of previous approaches in terms of pros and cons.

Algorithm	Research	Pros	Cons
Ontology	Our work	Constructed domain ontology converts unstructured text into structured data that can be directly applied for data analysis.	Creating ontology often relies heavily on manual efforts, prone to human errors and requiring considerable amount of time and cost.
Keyword-based	Lee et al. (2014), Kim (2012)	It is simple techniques for directly conveying information.	Correct results are retrieved only when the keyword matches exactly with the words or phrases within the database.

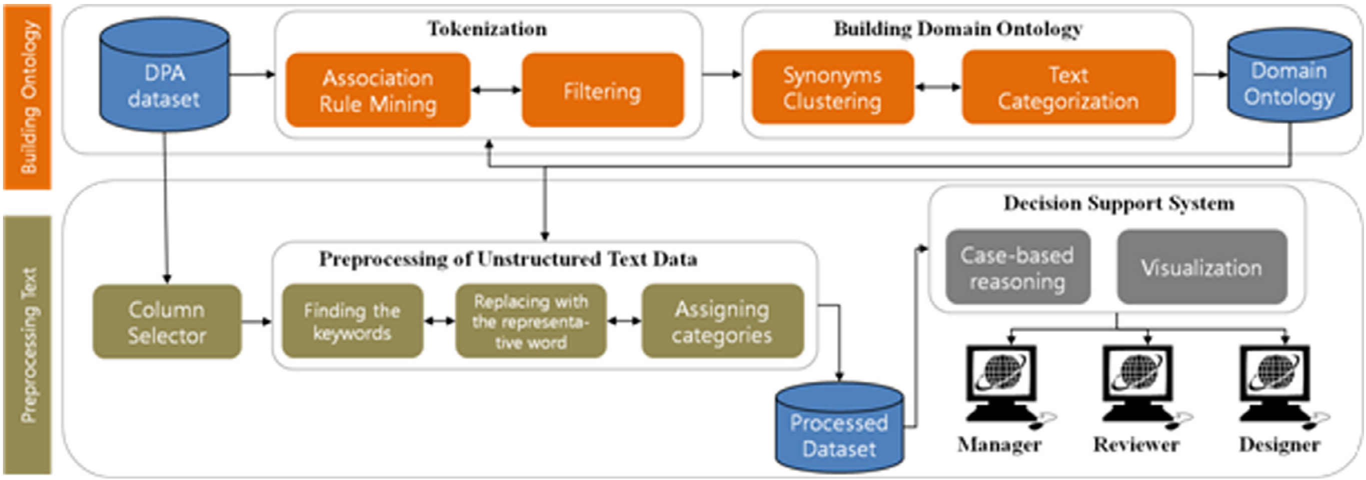


Fig. 2. Proposed knowledge discovery framework.

3. Knowledge extraction from unstructured text data

This section proposes a framework for extracting knowledge from unstructured text data. This framework will be applied on 114,793 verification reports stored in pre-assembly text database. The database contains 28 columns, including part type and design teams expressed as categorical variables, strings, codes or free-text. Amongst these columns, “problem” and “suggested solution” columns contain free-text, requiring ontology for knowledge discovery. Constructed domain ontology converts unstructured text into structured data that can be directly applied for data analysis. For other columns containing structured data, outlier detection and variable selection have been performed (Fig. 2).

3.1. Building domain ontology

Ontology is a collection of concepts based on semantic association. As its name implies, domain ontology denotes a knowledge system for a specific field, capturing meaningful concepts and semantic associations within the defined scope. Furthermore, it uses sub-class relationship and disjoint union to define relationship between concepts. Fig. 3 shows an example of domain ontology in an automobile industry. With “isSameAs” relationship, each concept is converted into a representative concept. For example, “part A1” is converted into its representative concept, “part A.” With “isA” relationship, each concept is matched to a corresponding higher-level concept. For example, “L-axis” is linked to its higher-level concept of “location.”

To enable more efficient data analysis, domain ontology removes unnecessary information in the free text data. As it is filled

Table 4
Examples of bigram and tri-gram and their corresponding support, confidence and lift values.

W1 W2 (W3)	Supp (ALL)	Supp (W1)	Supp (W2)	Supp (W3)	Supp (W1&W2)	Supp (W2&W3)	Conf (W1 → W2,W3)	Conf (W1,W2 → W3)	Lift (W1 → W2,W3)	Lift (W1,W2 → W3)
Brake tube	0.008	0.024	0.015	–	0.008	–	0.312	–	21.3	–
Rear door	0.007	0.051	0.032	–	0.007	–	0.142	–	4.4	–
Matching failure	0.007	0.027	0.031	–	0.007	–	0.273	–	8.7	–
Luggage side trim	0.003	0.009	0.033	0.022	0.005	0.004	0.332	0.586	84.4	26.5
Mounting hole unmatched	0.001	0.054	0.055	0.039	0.007	0.006	0.027	0.198	4.4	5.1

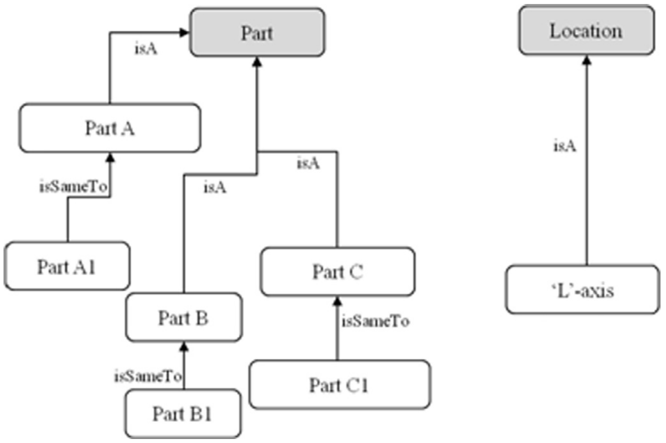


Fig. 3. The concept classification tree of automobile.

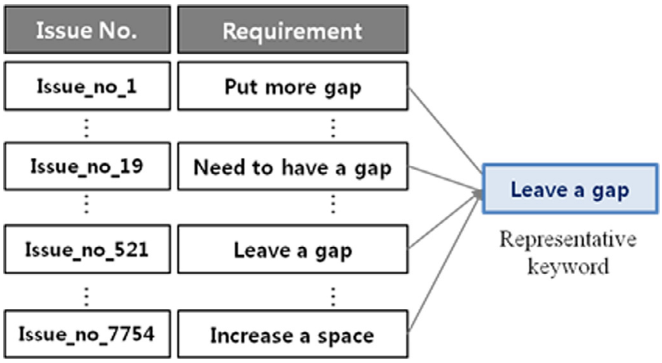


Fig. 4. Clustering concepts and assigning a representative keyword.

with technical terms and incorrect grammar, text preprocessing is necessary prior to constructing ontology. For preprocessing, tokenization, synonym clustering and categorization have been performed.

First, tokenization is applied to extract meaningful concepts within the data. If each concept is defined as an individual word, it fails to capture a concept composed of two or more words, consequently inhibiting effective data analysis. For example, concepts such as "brake pedal" or "weather strip" will lose their innate meanings if they are tokenized into "brake" and "pedal" or "weather" and "strip." In order to adjust for this problem, we use association rule mining on collocations to extract n-grams. Association rule mining (Agrawal, Srikant et al., 1994) has been widely used to explore co-occurrence of items within a single transaction. In our problem, each report is considered as single transaction, and each word as a single item, consequently generating various n-grams (Pecina, 2005). Through calculating confidence and lift of each n-gram, we denote meaningful n-grams as concepts. Confidence and lift can be calculated by Eq. (1) and dummyTXdummy-2. Support indicates a ratio of set X with respect to entire transaction, and confidence implies a probability of set Y occurring given set X. On the other hand, lift calculates a ratio of set X and Y co-occurring with respect to their independent occurrences. As shown in Table 4, we use these two measures for each n-gram to select appropriately tokened concepts from the reports.

$$conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X) \tag{1}$$

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{2}$$

Amongst these extracted concepts, semantically similar concepts are expressed in different forms as each reviewer has his or her own unique way of describing a problem (Rajpathak et al., 2012). As shown by the example in Fig. 4, leaving a gap between two components can be expressed in various different phrases such as "put more gap," "need to have a gap," "leave a gap" and "increase space." Consequently, semantically similar concepts are clustered, and represented by a single concept. As the reports are filled with technical terminologies, concepts are manually clustered by the domain experts. Table 5 lists some examples of clustered concepts.

Table 5
Examples of synonyms in the dictionary.

Representative concept	Synonyms
Mounting	MOUNTED, MOUNT, MT'G, MOUNTINGS, MOUNTING
Panel	PNL ASSY, PNL part, PANEL ASSY, PANEL, PNL
ASSY	ASS'Y, ASSEMBLY, assy, ASSY, ASY, compl, COMPL
Flange	FL, FLANGE
Pillar	PLR, FILLER, PILLAR, PLR part
Weather strip	W/STRIP, Weather strip
Shock absorber	SHK ABS, S/ABS, rear-sorber, absorber

As each representative concept is closely related to higher-level concepts, each concept is assigned to higher-level concepts during the categorization step. Due to their importance during the verification process, five categories (problem, cause, requirement, part, and location) are selected as higher-level concepts. For problem, cause and requirement, automatic assignment method is used to identify concepts belonging to these higher-level concepts. For the remaining two higher-level concepts, domain experts have manually identified the corresponding concepts. Table 6 provides a summary of each higher-level concept, their assignment methods and few examples. For automatic assignment method, a concept can be associated with different higher-level concepts depending on its context. Concepts belonging to higher-level concepts of problem and cause are especially susceptible to this issue. For example, "lack of gap" can be associated to a cause of an interference issue (e.g., "interference caused by lack of gap"), while it also can denote a problem (e.g., "lack of gap occurred because of the large size of components"). Consequently, frequent causal conjunction is used to capture the context of each concept and resolve this issue in automatic assignment method. Frequent causal conjunctions include terms such as "caused by," "due to," and "because of." Details of frequent causal conjunctions will be discussed in the following section.

Table 6
Categories of representative words.

Category	Description	Assignment	Examples
Problem	Words describing occurring problems, effect or phenomenon	Based on the predefined casual conjunctions, assign category automatically	Cannot assemble, Interference occurred
Cause	Words describing a cause or reason of the problem	Based on the predefined casual conjunctions, assign category automatically	Narrow, different shape, No holes
Solution	Words describing solutions to the problem	Words occurred in the Suggested Solution	Avoid interference, Check the data
Part	Words describing a name of a part or an assembly	Manually assigned by experts	Engine, Suspension, Door, Compressor
Location	Words describing a location where the part is assembled or the work is being done	Manually assigned by experts	Upper, Lower, Inner, Outer, Front, Rear

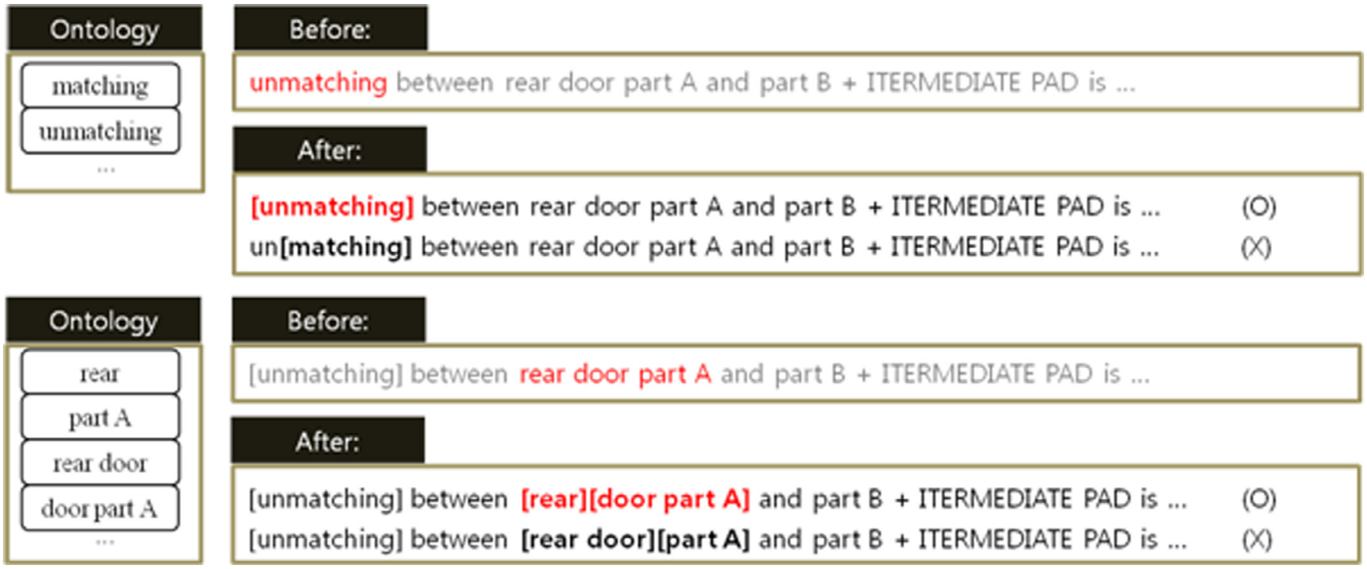

Ontology
matching
unmatching
...
Before:
unmatching between rear door part A and part B + ITERMEDIATE PAD is ...
After:
[unmatching] between rear door part A and part B + ITERMEDIATE PAD is ... (O)
un[matching] between rear door part A and part B + ITERMEDIATE PAD is ... (X)
Ontology
rear
part A
rear door
door part A
...
Before:
[unmatching] between rear door part A and part B + ITERMEDIATE PAD is ...
After:
[unmatching] between [rear][door part A] and part B + ITERMEDIATE PAD is ... (O)
[unmatching] between [rear door][part A] and part B + ITERMEDIATE PAD is ... (X)

Fig. 5. Finding the keywords using ontology.

As domain ontology generated from these processes can heavily influence the results of text preprocessing and thereby the quality of overall analysis, regular validation and re-training are required. Re-training can be accomplished by repeating the previous processes, through which new terminologies are added or errors in the ontology are corrected. Through regular updates, it can be continuously used for data analysis.

3.2. Ontology-driven text preprocessing

In order to select meaningful words or concepts from new reports, ontology-driven text preprocessing is crucial. During this process, a concept is extracted only if it exists within the generated ontology, and is replaced by its representative concept. Text preprocessing process includes finding keywords based on ontology, replacing input concepts by their representative concepts, and assigning them to higher-level concepts (categories).

To extract concepts, input text data's n-grams are looked up from the ontology. Although each word is usually 1 g, we start our search from longer n-grams as they have a higher chance of being more specific concepts. Furthermore, we start our search from the end of the input data as words in the beginning tend to be meaningless modifiers for more important concepts. Fig. 5 represents this process of extracting concepts. For the input word "unmatching," two possible concepts, "matching" and "unmatching," exist in the ontology. As "unmatching" is a longer string, it is selected as the extracted concept. For the input n-gram "rear door part A," possible concepts from the ontology include terms such as "rear door," "part A," "rear," and "door part A." Starting from the end of the input data, we select "rear door" and "part A" as extracted concepts.

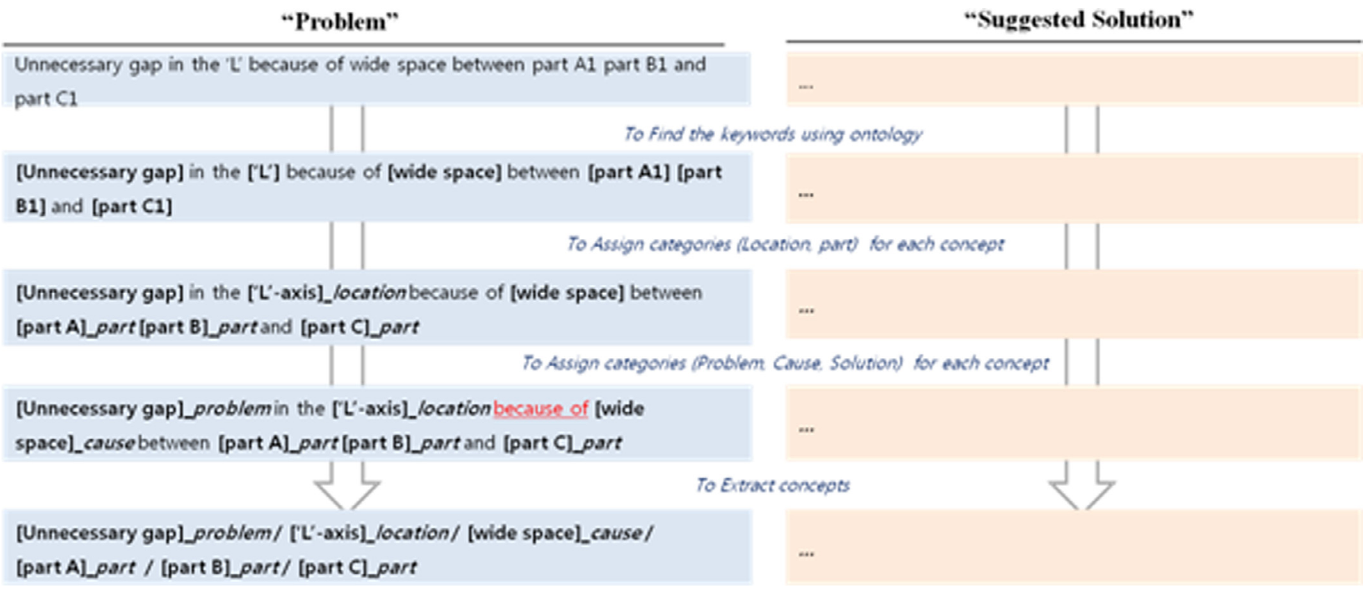
In order to account for different ways of expression amongst the engineers, extracted concepts are transformed into representative concepts. Using the ontology, we utilize heteronyms as representative concepts.

In order to assign each representative concept to higher-level concepts of location and part, we find its matching concept and its higher-level concept in the ontology.

To assign concepts to higher-level concepts of problem and cause, we determine the categorization by looking at the position of the concepts with respect to certain separators. On the other hand, the concepts generated from the requirement column in the database are all categorized as belonging to higher-level concept of requirement.

Finally, each concept's frequency and its higher-level concept categorization are acquired. Fig. 6 outlines the entire process of using ontology to preprocess new input records.

Table 7 shows few examples of final structured text data acquired from the verification reports after ontology-driven preprocessing. Using the domain ontology, we have systematically converted the text data from pre-assembly design database into a table containing each word's representative concept, frequency, and higher-level categorization. Fig. 7 shows sections of the domain ontology required for generating final structured text data in Table 7. Through "isSameAs" relationship, input words are converted into their representative concepts. Subsequently, they are categorized into higher-level concepts of part, cause, problem and solution through "isA," "isCauseInIssueNo," "isProblemInIssueNo," and "isSolutionInIssueNo" relationships. This structured text data,


"Problem"
"Suggested Solution"
To Find the keywords using ontology
To Assign categories (Location, part) for each concept
To Assign categories (Problem, Cause, Solution) for each concept
To Extract concepts
...

Unnecessary gap in the 'L' because of wide space between part A1 part B1 and part C1

[Unnecessary gap] in the ['L'] because of [wide space] between [part A1] [part B1] and [part C1]

[Unnecessary gap] in the ['L'-axis]_*location* because of [wide space] between [part A]_*part* [part B]_*part* and [part C]_*part*

[Unnecessary gap]_*problem* in the ['L'-axis]_*location* because of [wide space]_*cause* between [part A]_*part* [part B]_*part* and [part C]_*part*

[Unnecessary gap]_*problem* / ['L'-axis]_*location* / [wide space]_*cause* / [part A]_*part* / [part B]_*part* / [part C]_*part*

Fig. 6. Preprocessing with domain ontology.

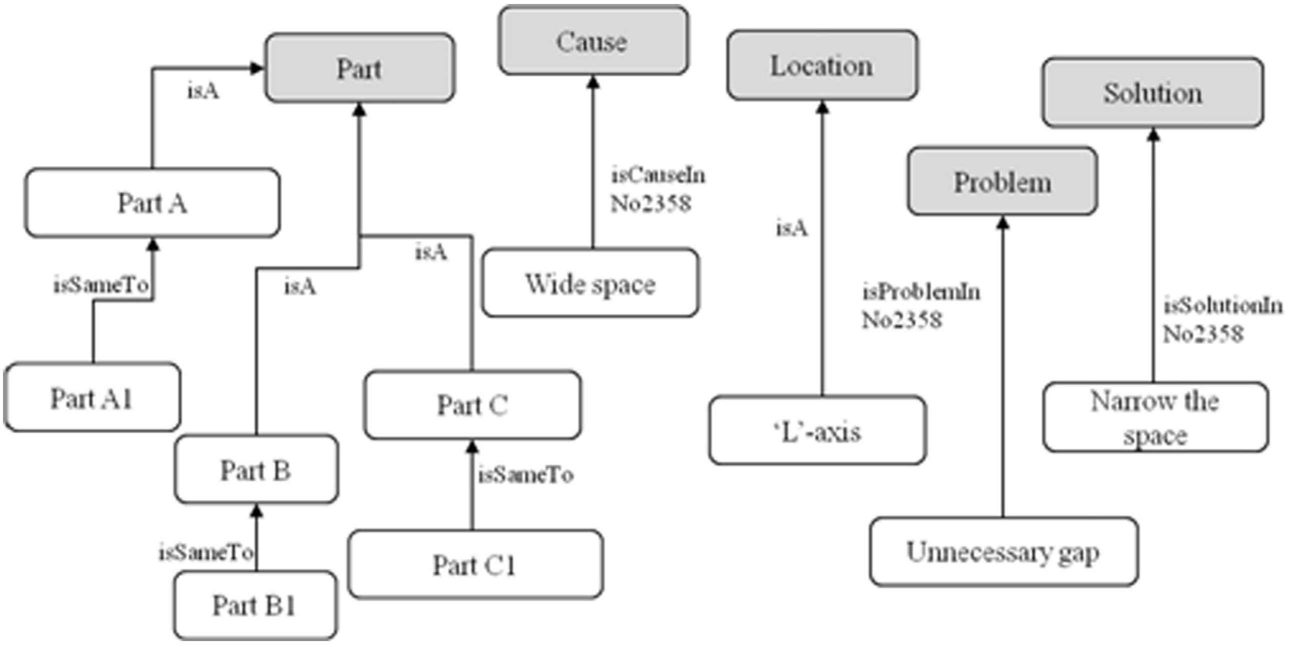


Fig. 7. Domain ontology from pre-assembly design database.

Table 7
Preprocessed text data.

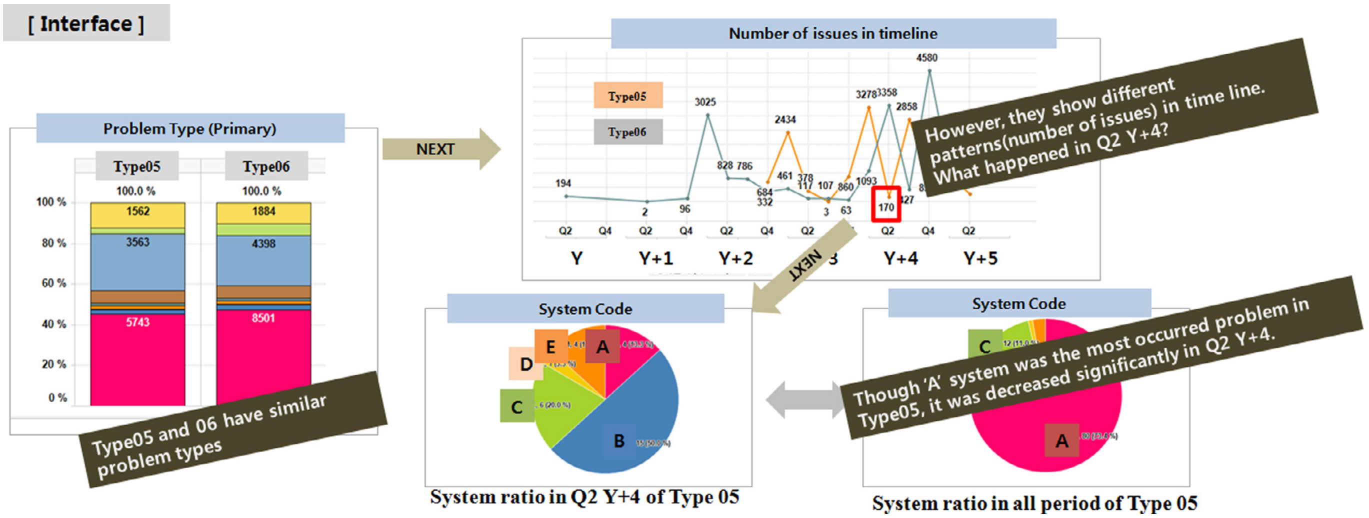
Issue No	Words	Freq	Category
...	...	...	...
No.2358	Part A	2	Part
No.2358	Part B	2	Part
No.2358	Part C	2	Part
No.2358	Wide space	1	Cause
No.2358	'L'-axis	1	Location
No.2358	Unnecessary gap	1	Problem
No.2358	Narrow the space	1	Solution
...	...	...	...

along with other structured data within the database, is used for knowledge discovery.

4. Visualization results

In the previous section, we have described the overall framework of extracting knowledge from unstructured database. With the domain ontology, we have successfully transformed free text into structured text data suitable for data analysis. In order to extract knowledge from the structured text data, we need to define types of knowledge required and methods for extracting required knowledge.

Main objective of this paper aims to improve the efficiency of the pre-assembly design verification process. Consequently, we define the types of knowledge and insights needed by three main contributors of the process: managers, reviewers, and designers. Considering each of the contributors' perspectives, we formulate specific datamining problems and analyze the data to retrieve the answers. To visualize our results, we have created a visual



Type05 and 06 have similar problem types

However, they show different patterns(number of issues) in time line. What happened in Q2 Y+4?

Though 'A' system was the most occurred problem in Type05, it was decreased significantly in Q2 Y+4.

System ratio in Q2 Y+4 of Type 05

System ratio in all period of Type 05

Fig. 8. Differences in raised issues between types of cars with developed interface.

interface that instantaneously calculates various measures for the user's queries.

In following three subsections, we will define each contributor's roles in improving the verification process. For each of three contributors, we will describe the type of knowledge required for fulfilling their roles. Using our visual interface, we will solve their problems in a drill down approach.

4.1. Knowledge discovery for managers

In design verification process, managers need holistic understanding of major problems in order to prepare response plans. Consequently, they want to identify which problems occur most commonly across all vehicles or how the issues differ amongst vehicles.

If they can discover any relationship between the assembly problems and the types of vehicles, they can formulate preventive plans to avoid similar problems in future product development, saving great amount of cost and time. Amongst various types of vehicles, we will investigate if the problems identified in Type05 cars are different from those of Type06 cars.

As depicted in the "Problem Type (Primary)" section in Fig. 8, the types of problems and their proportions are similar for both types of vehicles. If we observe the "Number of issues in timeline" section in Fig. 8, we notice that the number of issues for two vehicles differs as time progresses.

Therefore, we have decided to compare the patterns in the types of the issues. Although Type06 car's design verification has started from year Y, that of Type05 car has started 2 years later in year Y+2. Looking at their trends in the number of issues, we notice that they intersect with each other beginning in Q2 of year Y+4. In order to comprehend this observation, we will look at the system code ratio of these vehicles in Q2 of year Y+4.

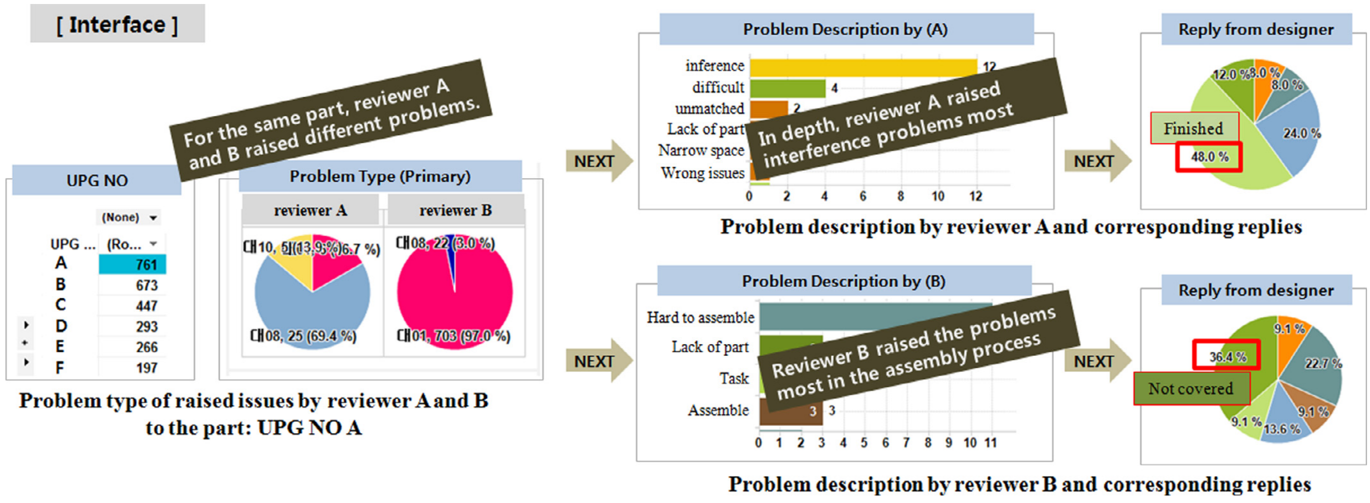
Throughout the entire period, the review team has been constantly identifying issues with system A. However, the frequency of this issue is significantly reduced during Q2 in year Y+4. Two possible reasons can explain this decrease. Either the designers have resolved the proposed issues or the reviewers might have started raising different issues, temporarily decreasing the frequency of the original problem. As the frequency of this issue is increased during Q3 in Y+4, later seems to be a more plausible explanation. Through verifying with the review team, we discover that the assigned reviewer has been changed at that time, and that the types of issues raised are heavily influenced by the assigned reviewer. Consequently, we conclude that it is important to understand the patterns of the issues posed by each reviewer. Furthermore, we observe that the managers need to understand both the reviewers and the designers in order to comprehend the entire process.

4.2. Knowledge discovery for reviewers

Each reviewer is responsible for evaluating a specific design issue. However, the contents of the reports might vary even amongst the reviewers with same responsibilities. As clear and detailed review criteria do not exist, the reports are heavily influenced by the reviewer's personal tendencies. Furthermore, these reports are not shared between the reviewers, generating redundant evaluations. In order to remedy this inefficiency, we can compare each reviewer's personal tendencies during their evaluation to adjust for his or her bias. To adjust for these tendencies, the review teams want to discover the trends in the review teams' evaluation for each type of vehicles, responses of design teams with respect to each review team, and effect of change in review teams' personnel on the reviews. Amongst these various questions, we will confirm whether certain reviewers raise different issues compared to other reviewers.

Fig. 9 describes a visual interface that reveals how two reviewers have evaluated the design for an identical component. Amongst various issues posed by the reviewers, we select the most frequent issue regarding part UPG NO A for further analysis. Observing the types of the issues in the "Problem Type (Primary)" section of Fig. 9, we discover that Reviewer A often raises problem type 08, while Reviewer B frequently suggests problem type 01. This difference confirms that each reviewer's subjective tendencies do influence the evaluation result.

Looking at the corresponding text data, we discover that Reviewer A frequently raises "interference" related issues, while Reviewer B poses issues regarding "assembly process." As their primary problem type differs, the contents of their reports are inevitably dissimilar. Looking at the designers' response to two reviewers' reports, we discover that 48% of Reviewer A's reports are replied as "finished," while 36.4% of Reviewer B's reports are replied as "not covered." Consequently, we can conclude that the reports filed by Reviewer A have a higher chance of being accepted by the designers. Furthermore, we can investigate for reasons why Reviewer B's evaluation is often ignored by the designers. By using our visual interface, we can objectively compare differences



For the same part, reviewer A and B raised different problems.

In depth, reviewer A raised interference problems most

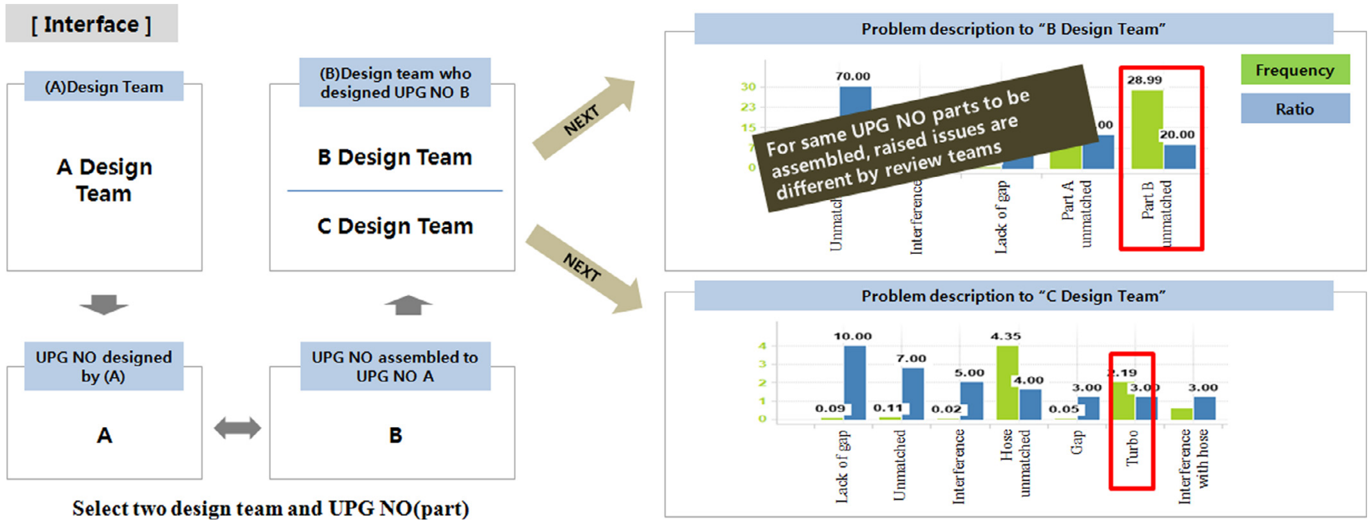
Reviewer B raised the problems most in the assembly process

Problem type of raised issues by reviewer A and B to the part: UPG NO A

Problem description by reviewer A and corresponding replies

Problem description by reviewer B and corresponding replies

Fig. 9. Reviewers raise different issues for the same part.



For same UPG NO parts to be assembled, raised issues are different by review teams

Select two design team and UPG NO(part)

Fig. 10. Problems raised for the part designed by two teams.

between the reviewers, enabling us to adjust for their evaluation tendencies and bias.

4.3. Knowledge discovery for designers

During pre-assembly process in automobile manufacturing, most issues occur when two components are assembled together. When problems are detected, the designers responsible for creating the problematic components are also notified. Similar to Section 4.2, types of issues raised can be different even for an identical component due to individual designer's preference in design. Consequently, a designer can avoid assembly issues if one can incorporate the preference of the other designers, responsible for creating different components for the assembly.

Furthermore, a designer might constantly receive redundant feedbacks due to one's carelessness. If one is willing to understand which aspects of his or her design preferences have caused these redundant feedbacks, one can easily remedy this issue. In order to explore different design preferences amongst the designers, we analyze the types of issues received by each designer.

As the number of data decreases significantly at an individual designer level, we have decided to compare the issues received by the design teams. As shown in Fig. 10, design team A receives various types of issues when its component "UPG NO A" is assembled with component "UPG NO B." Two different design teams, B and C, have created "UPG NO B." Observing the issues received by design team B and C as shown in Fig. 10, we notice that these two teams receive both similar and different types of issues. Amongst these issues, both teams most frequently receive issues related to "unmatched part A." However, the second most frequent issues are different for each team. While design team B receives issues with "unmatched part B," team C receives problems related to "turbo." Beside these issues, problems related to "gap" and "interference" have also been commonly raised for both teams.

As shown by above, we confirm that the types of issues raised can differ depending on the collaborating design teams. Using our graphical user interface, the design teams are now able to answer other meaningful questions such as change in types of issues received depending on a review team or a type of vehicle.

5. Conclusion

During digital pre-assembly process, problems and requirements for the components are recorded in part verification reports. Despite a huge number of reports, it is difficult to understand causes of these issues as the reports are recorded in free-text format. This paper proposes a knowledge discovery framework for pre-assembly review database. Up to now, it is the only paper that applies text mining techniques to digital pre-assembly process. In order to utilize the text data, we first create domain ontology. Using this ontology, we subsequently preprocess the reports collected from the pre-assembly process. This system of converting free text into structured text data can be continuously used through regular updates.

During visualization, we combine our preprocessed text data along with other structured data to extract meaningful insights. In this paper, we define three real business problems faced by managers, reviewers and designers. In order to answer their business problems, we apply text mining techniques to extract knowledge from text data produced in the pre-assembly process. With friendly graphical interface, we provide clear visualization for our extracted insights.

Major contribution of this paper is that it applies text mining techniques to the manufacturing industry. Within automobile manufacturing, previous studies have analyzed data for quality improvement. However, no previous studies have applied text mining on data from the pre-assembly process. Furthermore, the text data analyzed in this paper contains both English and Korean, which is unprecedented even in other industries.

Furthermore, this paper proposes a method for constructing ontology for automobile design terminologies. As text data in the pre-assembly process is filled with technical terms and acronyms, domain ontology is crucial for data analysis in the automobile industry. Furthermore, we create a system for preprocessing new text data using the generated ontology.

We have also created a clear graphical interface that can visualize the outcomes of text mining. This tool can be expanded to solve various other business problems not dealt in this paper.

However, there is a limitation to our proposed approach. In this paper, semantically similar terms have been manually clustered, requiring huge amount of time and human efforts. If we apply topic modeling techniques such as LDA, we can automate our domain ontology building process. However, it is difficult to visualize the results of LDA, suggesting possible future research topic.

As future works, we will improve the quality of our ontology by regularly preprocessing new text data and updating it. Through enriched ontology, we can acquire more reliable results from data analysis. Furthermore, we can combine our data with other data from different sources to extract more meaningful business insights. This paper uses data from the digital pre-assembly process to resolve design issues of components. If we can incorporate data generated from customers' vehicles or voice of customers, we can provide more robust design guidelines, and offer more significant business insights for quality improvement.

Acknowledgments

This work was supported by the BK21 Plus Program(Center for Sustainable and Innovative Industrial Systems, Dept. of Industrial Engineering, Seoul National University) funded by the Ministry of Education, Korea (No. 21A20130012638), the National Research Foundation of Korea(NRF) grant funded bythe Korea government(MSIP) (No. 2011-0030814), and the Institute for Industrial Systems Innovation of SNU.

References

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB: 1215* (pp. 487–499).

Baba, Y., & Nobeoka, K. (1998). Towards knowledge-based product development: The 3-d cad model of knowledge creation. *Research Policy, 26*(6), 643–659.

Chakrabarty, S., Chougule, R., & Lesperance, R. M. (2009). Ontology-guided knowledge retrieval in an automobile assembly environment. *The International Journal of Advanced Manufacturing Technology, 44*(11–12), 1237–1249.

Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2009). Data mining in manufacturing: A review based on the kind of knowledge. *Journal of Intelligent Manufacturing, 20*(5), 501–521.

Chougule, R., Rajpathak, D., & Bandyopadhyay, P. (2011). An integrated framework for effective service and repair in the automotive domain: An application of association mining and case-based-reasoning. *Computers in Industry, 62*(7), 742–754.

Huang, Y., McMurran, R., Dhadyalla, G., & Jones, R. P. (2008). Probability based vehicle fault diagnosis: Bayesian network method. *Journal of Intelligent Manufacturing, 19*(3), 301–311.

Khare, V. R., & Chougule, R. (2012). Decision support for improved service effectiveness using domain aware text mining. *Knowledge-Based Systems, 33*, 29–40.

Kietz, J.-U., Volz, R., & Maedche, A. (2000). Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd workshop on learning language in logic and the 4th conference on computational natural language learning-volume 7* (pp. 167–175). Association for Computational Linguistics.

Kim, B. (2012). *Analysis on engineers process event log in semiconductor processing with text mining technique*. Seoul National University.

Lee, S.-k., Kim, B., Huh, M., Park, J., Kang, S., Cho, S., et al. (2014). Knowledge discovery in inspection reports of marine structures. *Expert Systems with Applications, 41*(4), 1153–1167.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation & Computational Linguistics, 11*(1–2), 22–31.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL student research workshop* (pp. 13–18). Association for Computational Linguistics.

Rajpathak, D., Siva Subramania, H., & Bandyopadhyay, P. (2012). Ontology-driven data collection and validation framework for the diagnosis of vehicle health management. *International Journal of Computer Integrated Manufacturing, 25*(9), 774–789.

Rajpathak, D. G., & Singh, S. (2014). An ontology-based text mining method to develop d-matrix from unstructured text. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44*(7), 966–977.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning, 34*(1), 233–272.

Yao, Z., & Ze-wen, C. (2011). Research on the construction and filter method of stop-word list in text preprocessing. In *Intelligent computation technology and automation (ICICTA), 2011 international conference on: 1* (pp. 217–221). IEEE.