The Assistant Analysis of Defect Cause based on Knowledge Graph and Texts

Jiayuan Xie

School of Guangdong University of Technology, Guangzhou 510000, China

Jiayuan.xie@qq.com

Abstract

With the development of the Internet, more and more problem sheets in the manufacturing line are generated by technologists, which are mainly composed of two types of texts: defect description and cause description. Analyzing the causes of the specific defect can guide technologists to deal with and solve the defect. Traditional information retrieval methods only focus on the keywords appearing in the defect description texts and return the cause description texts which contain the same keywords. However, it will pose two difficulties. The amount of cause description texts is so big that it is hard for technologists to mining useful information in a short time. What's more, current approaches lack domain background knowledge. To address these two challenges, the paper introduces an approach to help technologists analyse the causes of defects in a cost-benefit way, which is a semi-automatic approach. First, we apply a text clustering method to group similar cause description texts to get different causes of a specific defect rather than amount same cause description texts in a cost-benefit way. Then a domain knowledge graph is constructed which contains abundant domain information, to link the relations between the cause description texts. Experiments demonstrate our approach's ability in helping technologists analyse the causes of specific defect more efficiently and effectively.

Keywords

Assistant Analysis, Knowledge Graph and Texts.

1. Introduction

With the development of the Internet, more and more problem sheets in the manufacturing line are generated by technologists, which are mainly composed of two types of texts: defect description and cause description. For example, when a defect "power tube open wilding" comes, a problem sheet is made up of two parts: the defect description text "on 2012.3.4, power tube open wilding" and the cause description text "power tube's shape is changed". Analyzing cause description texts of defects can help and guide technologists in the manufacturing line to know causes for defects and thus they can know how to deal with defects.

Traditional information retrieval methods only focus on the keywords appearing in the defect description texts and return the cause description texts which contain the same keywords. As the example mentioned above, traditional information retrieval method will return the cause text "power tube's shape is changed" to technologists, which contain the keyword "power tube", rather than text "pin's shape is changed".

However, it will pose two difficulties. The amount of cause description texts is so big that it is hard for technologists to mining useful information in a short time. According to our observation and analysis of cause description texts, most of them are duplicated. For example, when the defect "power tube open wilding" comes, there are thousands of cause description texts which contain the keyword "power tube" will be returned to technologists. As a result, technologists will spend most of time in doing repetitive work, such as browsing thousands of repetitive texts and find out the duplicates one by one, which will cost a lot of time and money.

What's more, current approach lacks domain background knowledge. Let's take the example mentioned above into consideration, the cause description text "pin's shape is changed" is also the description of the defect "power tube open wilding". However, as it does not contain the keyword "power tube", it won be return to technologists. On the other hand, if machines do not know that "pin" is related to "power tube", only according to their surface meaning, machines can not understand the implicit meaning of the cause description texts, such as the relation between "pin" and "power tube".

To address these two challenges, the paper introduces an approach to help technologists analyse the causes of defects in a cost-benefit way, which is a semi-automatic approach. First, we apply a text clustering method Organization of the Text

2. Cause Description Texts Analysis.

Table 1: A SAMPLE OF CAUSE DESCRIPTION TEXTS OF DEFECT ”POWER TUBE OPEN

Defect Description Text	Cause Description Text
Defecta	Power tube deformation
Defecta	Change power tube’s shape
Defecta	The shape of power tube is changed

There are thousands of problem sheets generated by technologists in the manufacturing line. A problem sheet is made up of two parts: the defect

description text and the cause description text. Analysing the cause of the defect can help technologists understand and deal with the defect. Traditional information retrieval methods only focus on the keywords in the defect description texts and then return the cause description texts which contain the same keywords to technologists.

Table 2:THE NUMBER OF CAUSE DESCRIPTION TEXTS OF VARIOUS DEFECTS

Defecta	The Number of Cause Description Texts of The Defect
Defectb	2400
Defectc	3000

For example, we can see a sample of defect description texts and cause description texts about the defect "power tube open welding" in table1. When the defect "power tube open welding" comes, current approach will use "power tube" as a keyword and then return cause description texts which contain the same keyword such as "power tube deformation", "change power tube's shape" and "the shape of power tube is changed" to technologists.

Current information retrieval approach will pose two difficulties. First, Table 2 description texts is so big that it's a time-consuming and money-consuming process for technologists to mining useful information from them. As a result, how to help and guide technologists understand and analyse the amount cause description texts and then mining useful information from them is a big challenge.

After analysing of the cause description texts, we find that there are many cause description texts share the same meaning in different ways. For example, in table 2 ,"power tube deformation", "the shape of power tube is changed" and "change power tube's shape" express the same meaning "power tube's shape in changed" in three ways. According to our observation, we apply K-means [5], a text clustering method, to group similar cause description into ten clusters.

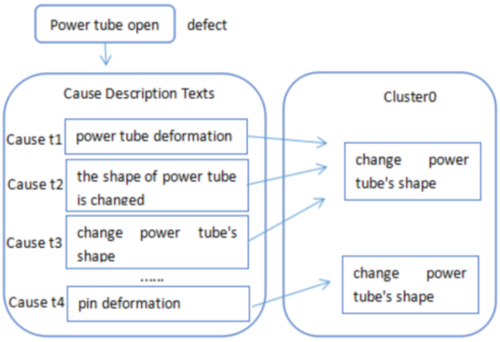

Fig. 1. The final cause description text

Figure 1 shows the final results of cause description texts after applying K-means in cause description texts of the defect "Power tube open welding", "power tube deformation", "the shape of power tube is changed" and "change power tube's shape" will be grouped into one cluster and with the help of experts, texts in the cluster will be summarized as "power tube deformation", which is a semi-automatic approach.

K-means can group similar texts (those cause description texts which share the same meaning) into ten clusters, which is a automatic process. Experts will spend less time in doing repetitive work, such as browsing thousands of repetitive texts and find out the duplicates one by one, which is a time-consuming and money-consuming process. On the other hand, experts can pay more attention to the analysis of cause description texts and summarize them in a more efficient and effective way.

3. Domain Knowledge Graph.

The second difficulty posed by current approach to analyse the causes of defects is that they lack domain knowledge. For example, when the defect "power tube open welding" comes, if we do not know "pin" is a component of "power tube", we can not realize that "pin" may be the cause of the defect even though they do not have the same surface meaning.

There are many knowledge graphs, such as WordNet [1], a lexical knowledge base for English language, and open domain knowledge graph DBpedia [2], YAGO [3] , and Probase [4], which focus on general knowledge. However, none of existing knowledge graph are designed for professional domain. Therefore, we study how to build a specific domain knowledge graph by making use of the cause description texts in the problem sheets in the manufacturing line.

After analysing the clustering results in the previous step, we only focus on the components in the cause description texts, derived from the following observation. We find that the similar cause description texts in one cluster are always about a kind of component. As the example mentioned above, "power tube" is a component and the cluster which the cause description texts are grouped in are mostly related to "power tube". What's more, according to our observation and analysis of the final cause description texts, it is surprising that most of clusters which contain similar cause description texts are always related to some components, such as "power tube", "pin", and "positioning".

Knowledge graphs can help technologists find the relations between two entities, such as two components. On the other hand, knowledge graphs can help technologists find the relations between two components in different cause description texts although they do not have the same surface meaning, such as "pin" and "power tube". As a result, mining the components from final cause description texts can help technologists analyse the causes in a more

3.1 Graph Definition

According to our observation, we propose to construct a domain knowledge graph about defects, components and their relations. Generally speaking, there are three kinds of nodes in our knowledge: defect category, defect and component. Each node of the knowledge graph belongs to one node category.

1) Defect Category: A node in this level denotes a category which a defect belongs to. For example, "fruit" is a category, "apple", "grape" and "orange" are those instances belong to the category. In our domain knowledge graph, defect category represent what kind of problem the defect belongs to, such as "process problem", "device problem" and "human issues".

2) Defect: This level contains all defects in the cause description texts of problem sheets, which are the most important part of our domain knowledge graph. Each node represents a defect such as "power tube open welding" and "printing less tin". Different defects belong to different categories and thus the "belong to" relation will link a defect and it's category.

3) Component:A defect is usually related to a component and thus the relation "related to" will link a defect and a component which is related to the defect. There are many node which represent components in our domain knowledge graph, such as "power tube", "pin". What's more, a component also can be a component of another-one component. Let's take "power tube" as an example, "pin" is a component, and it is also is a component of "power tube". As a result, the relation "component of" will link the relations between two component.

3.2 Knowledge Graph Building

We propose to construct a domain knowledge graph based on cause description texts. As mentioned above, there are three kinds of nodes in our domain knowledge graph. Figure 2 shows a sample of our domain knowledge graph. We describe the details in constructing each kind of nodes in our domain knowledge graph.

1) Defect Category and Defect: We extract defects and defect categories from cause description texts and domain dictionaries. There are 663 defects and they belong to 23 defect categories in our domain knowledge graph. What's more, we identify "belong to" relation from domain dictionaries based on rules. For example, from "process problem contains the problem of power tube open welding" we can obtain that "power tube open welding" is a kind of defect and it belongs to "process problem", which is a kind of defect category.

2) Component: As mentioned above, the clusters obtained from applying K-means to the cause description texts are almost related to a component. Inspired by idea from \cite{aspect}, defect is an entity and the component it is related to is the entity's component, which describe some aspects about the entity. What's more, component can be a sub-component of another component. For example, we know "power tube open welding" is a defect, and it is related to many components, such as "power tube" and "pin". "Power tube" and "pin" are two kinds of components and "pin" is a sub-component of "power tube". There are many state-of-art methods to extract aspect for opining mining, such as [7], [8] and [9]. In this paper, we use a automatic approach to extract components from cause description texts. We extract some frequent noun phrases by using syntactic parsing, [10] and based on rules, which we regard them as components in our domain knowledge graph. These extracted components are regarded as nodes in our knowledge graph.

4. Experiment

We conduct comprehensive experiments on real-world dataset to demonstrate the ability of our approach in helping technologists in the manufacturing line analyse the defect causes in a efficient and effective way. Table 3 shows the comparison of the origin number of cause description texts and the number of final cause description texts by using the semi-automatic approach. The semi-automatic approach can help decrease the number of duplicate texts and experts's analysis of the clustering results will increase the precision and the quality of the final results.

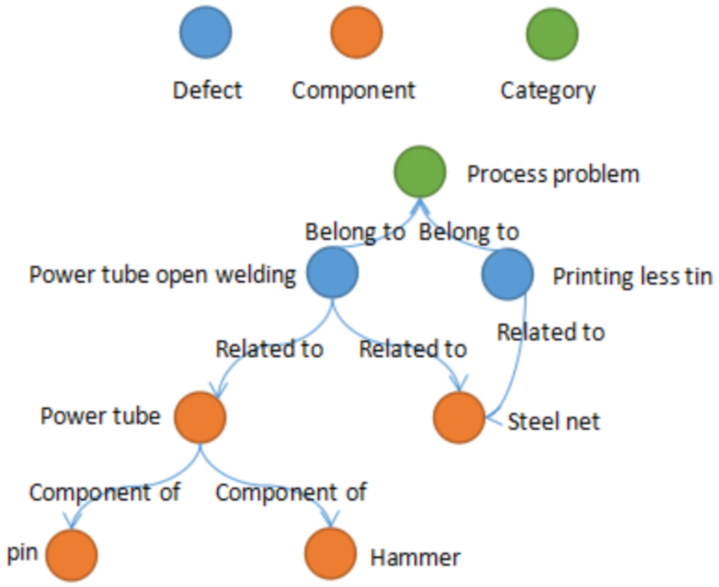

Fig. 2. A sample of knowledge graph

Tabel 3: STATISTIC OF THE CAUSE DESCRIPTION TEXTS

Defect	Origin Number	Clusters	Average number
defecta	2400	10	2-5
defectb	3000	10	3-6
defectc	1800	10	3-5

In order to evaluate effectiveness of our constructed domain knowledge graph, we perform a case study to demonstrate the effectiveness of our domain knowledge graph. Figure 2 shows a sample of our domain knowledge graph. When the defect "power tube open welding" comes, traditional information retrieval methods can only return those cause description texts which contain "power tube" to technologists. With the help of our domain knowledge, technologists can also get those cause description texts which contain "pin" as "pin" is a component of "power tube" in our domain knowledge graph.

5. Conclusion

With the development of the Internet, more and more problem sheets are generated by technologists, which are made up of two kinds of texts. Analysis of defect causes can help and guide technologists process and deal with defects. Traditional information retrieval methods can not work well as there are thousands of cause description texts and it's hard for technologists to mining information from them. What's more, current approach lack domain knowledge. To address this two challenge, this paper introduces a semi-automatic approach to help technologist analyse the defect causes, which firstly apply K-means to cause description texts and then construct a domain knowledge graph.

References

[1] George A. Miller, ``WORDNET: A Lexical Database for English," Commun. ACM. , vol. 38 , pp. 39--41, 1992.

[2] ren Auer and Christian Bizer and Georgi Kobilarov and Jens Lehmann and Richard Cyganiak and Zachary G. Ives, DBpedia: A Nucleus for a Web of Open Data, ISWC/ASWC, 2007.

[3] Fabian M. Suchanek and Gjergji Kasneci and Gerhard Weikum, Yago: a core of semantic knowledge, WWW, 2007.

[4] Wentao Wu and Hongsong Li and Haixun Wang and Kenny Q. Zhu, Probase: a probabilistic taxonomy for text understanding, SIGMOD, 2012.

[5] Bottou, Leon and Bengio, Yoshua and others, Convergence properties of the k-means algorithms, Advances in neural information processing systems, MORGAN KAUFMANN PUBLISHERS, 1995, pp. 585--592.

[6] Zhiyuan Chen, Arjun Mukherjee and Bing Liu, Aspect Extraction with Automated Prior Knowledge Learning. ACL, 2014.

[7] Ruidan He and Wee Sun Lee and Hwee Tou Ng and Daniel Dahlmeier, An Unsupervised Neural Attention Model for Aspect Extraction. ACL, 2017.

[8] Lei Shu and Hu Xu and Bing Liu, Lifelong Learning CRF for Supervised Aspect Extraction. ACL, 2017.

[9] Yichun Yin and Yangqiu Song and Ming Zhang, Document-Level Multi-Aspect Sentiment Classification as Machine Comprehension. EMNLP, 2017.

[10] Zhenghua Li, Min Zhang, Wanxiang Che, Ting Liu, Wenliang Chen, and Haizhou Li, Joint Models for Chinese POS Tagging and Dependency Parsing. EMNLP, 2011.07, pp. 1180-1191. Edinburgh, Scotland, UK.