SAS® GLOBAL FORUM 2020 #SASGF

5175-2020

Developing an Automotive Safety Ontology through Concept Map and Text Analytics

Sadikshya Basnet, Oakland University

ABSTRACT

Vehicle safety is an important area in the automotive sector; however, there is no standard terminology that all the stakeholders can use. Ontologies have long been argued as one approach for capturing and representing domain knowledge. Ontologies define the terminology of a domain by specifying the relevant hierarchical concepts and their relationships. Ontology development is an expensive and time-consuming process. This paper proposes a concept map-based approach for automotive safety ontology development by first semi-automatically creating a detailed level entities/concepts as a keyword list by applying natural language processing, including word dependency and POS tagging. Specifically, SAS Text Miner 15.1 will be used for analyzing the customer complaint dataset published by NHTSA. The ontology development workflow will include standard text mining nodes such as Import, Parsing, Filter, Topic, and Cluster for processing the customer complaint text and deriving safety related terms and relationships. This is then used to extract appropriate entities/concepts and develop the concept map and eventually an ontology for the automotive safety domain. Having a unified ontology will greatly help in minimizing the miscommunication between various stakeholders and ensure that the designers, suppliers, manufacturers, dealers, and repair shops are all on the same page with respect to automotive safety related issues. The intended audience for this presentation is SAS users who are working in the area of text analytics and automotive safety professionals.

INTRODUCTION

Ontologies define the terminology of a domain by specifying the relevant hierarchical concepts and their relationships. They can easily include tens or hundreds of thousands of concepts and are both expensive and time-consuming to develop. Ontology creation is the process of automatically or semi-automatically constructing ontologies based on textual domain descriptions. The assumption is that the domain text reflects the terminology that should go into an ontology, and that appropriate linguistic and statistical methods should be able to extract the appropriate concept candidates and their relationships from these texts.

Generating concept maps through text analytics has received increased attention and creating concept maps in a semi-automated manner is becoming feasible through several tools and templates. Thus, creating a concept map from existing documents and other knowledge sources in a domain and using it as a starting point and transforming it into an ontology is quite appealing. Hence, the objectives of this paper are to develop a) an approach for creating concept maps from a set of domain documents, b) transform a concept map into a corresponding ontology, and c) demonstrate the feasibility of the approach using a case study.

RELATED WORK

Some ontology workbenches exist that support the creation of ontologies. The JATKE workbench is implemented as a plug-in to the Protégé ontology editor and helps users develop ontologies in Protégé [1]. OntoLT is another Protégé plug-in that transforms linguistically annotated entities into concepts and individuals in ontologies [2]. Text2Onto is an advanced ontology workbench that makes use of both Lucene and GATE [3] to produce many of the same ranked candidate lists [4]. Text2Onto also has some additional features for ontology maintenance and incremental ontology updating. The OntoLearn workbench from Navigli and Velardi [5] concentrates on word sense disambiguation and makes use of the WordNet lexicon. A common aspect of all these workbenches is that they consider the ontology creation process a chain of static analysis components. The user does not have the flexibility to adapt the analysis to his or her preferences, the nature of the document collection, or aspects of the domain itself. Furthermore, it is assumed that the user is familiar with ontology editors and can afterwards refine the generated results manually.

PROPOSED APPROACH

The Dataset

The dataset used in this research is the National Highway Traffic Safety Administration (NHTSA) public data[1]. The NHTSA is an agency of the Executive Branch of the U.S. government, part of the Department of Transportation. The data is provided by Office of Defects Investigation (ODI) in NHTSA. The whole database dump contains about 1.5 million of vehicle safety complaints records since 1995. The data resource is consumers' complaint about the vehicle incidents. Each record includes a unique ID (ID), manufacturer' name (MFR_NAME), vehicle/equipment make (MAKE), vehicle/equipment model (MODEL), model year (YEAR), date of incident (FAIL DATE), specific component's description (COMPDESC), detailed information about consumer's vehicle (e.g., VIN number), and the content of the complaint (CDESCR). In this research, we extracted the information from the content of the complaint to construct knowledge map then mapping the results to the ontology. An example record in the dataset is shown in Table 1. Some of the columns in the dataset are omitted here for space.

Table 1. An Example Record in the NHTSA Complaint Dataset

ID	MFR_NAME	MAKE	MODEL	YEAR	FAIL DATE	COMPDESC	VIN	CDESCR
1000051	Ford Motor Company	FORD	FUSION	2010	20130718	VEHICLE SPEED CONTROL	3FAHP0HA1AR	Vehicle keeps shutting off while driving. First happened on July 18th, second July 19th. Just started again on July 31 & continues. I was told it's my throttle.

[1] Dataset from NHTSA link: https://www-odi.nhtsa.dot.gov/downloads/

Concept Extracting and Mapping

Extracting. From the content of the consumers' complaints, we identified that the common pattern of the complaints is that the consumers first described the situation or the contextual information about the incidents, then the consumers used their own terms and vocabulary to describe how exact the incidents happened, and the out-comes. Fortunately, for each complaint record, there is a higher-level component as one automobile property identified, such as engine, power train, electrical system, etc. However, the lower level components that involved in the incidents can only be found from the content of each complaint. And these components are usually not represented by using a formal term or followed the terminology used by manufacturer.

The first task of our approach is to extract these automotive components from the content of the complaint. We applied the Stanford NLP package to perform tokenization, word dependency, and POS tag parsing, to identify nouns as entities of components. Since the dataset has several different manufacturers, models, and years, to keep the components consistent, we randomly chose a single manufacturer to build a training subset of complaints, then to extract component keywords.

From about 1.5 million complaint records, we extract 10,000 records by matching the chosen manufacturer. After the POS tag parsing, all of the nouns (i.e., with POS tags as NN, NNP, NNS, NNPS) are identified and counted. There are 8,393 unique nouns, from which we manually picked the top 300 nouns (ranked by number of term frequency count descending) that related to automotive components. The component keyword list consists of these 300 nouns. Moreover, to capture the component with more than one term, we performed a 2-gram nouns phrase matching if both terms in the 2-gram are matched the list of 300 nouns.

Mapping. In the complaint dataset, each record has a high-level component cate-gory that identified by NHTSA. We use this category as a higher-level concept in the ontology to be mapped with the components extracted from the complaint content. The new relationships are created by this mapping process. Table 2 shows the relationships from an example complaint record.

Table 2. The Relationships from an Example Complaint

COMPDESC	Content of Complaint	Extracted Entities/Components	
		Single Term	2-Gram
Engine	The wrench light comes on and the car loses its acceleration completely. As I pull over the side of the road and brake, the car will start shaking. The only way to fix it is to turn the car off and back on again. The wrench light will disappear after that and it'll start driving normally again. First time this happened I was on the interstate with my three-year-old in the car with me. One unhappy momma!!!! The code it read p2111 which I have been told it is a defect in the throttle body. Hope they are right.	Wrench Light Brake Code Throttle Body	Wrench Light Throttle Body

From Table 2, we can see that the Engine entity as a high-level component has the relationships with wrench light, throttle body, brake, and code (p2111) in this incident. These relationships are used to construct the concept map to further develop the ontology over time.

Overall Process using SAS Text Miner

We implement our proposed approach in SAS Text Miner 15.1 with the following steps.

Complaint dataset loader. Import the raw dataset and remove the duplicated records.

Preprocessor and NLP parsing. The NLP parses the complaint content for tokenization, POS tagging, and creating bag of words.

Keyword creating. The single terms are aggregated for manual process to filter out the terms are not related to automotive component.

Keyword loader. Load the keyword list from previous component.

Single term/2-Gram extractor. Using the keyword list to match single keyword or 2-Gram keywords. The results can be further aggregated at different levels from the complaint dataset loader.

Initial Results

Aggregation. Table 2 only shows one complaint record's result. The results from multiple complaints can be aggregated at different level. For example, the single term and 2-Gram results can be aggregated at COMPDESC (high level component) level, which can be engine, power train, etc. Other possible levels can be any columns in NHTSA dataset. Moreover, the results can be aggregated by more than one level at the same time, such as model and year. Table 3 shows the top 10 2-Grams aggregated result for different COMPDESC by descending order from 10,000 complaint records related to manufacturer Ford Motor Company. Table 4 shows the top 10 2-Grams aggregated result for different model and year by descending order from 10,000 complaint records related to manufacturer Ford Motor Company.

Table 3. 2-Gram Aggregated Result based on COMPDESC

COMPDESC	2-Gram	Count
Vehicle Speed Control	Throttle body	449
Steering	Power steering	395
Power Train	Throttle body	250
Steering	Steering wheel	229
Engine	Throttle body	168
Vehicle Speed Control	Gas pedal	142
Power Train	Wrench light	138
Engine	Engine light	131
Fuel/Propulsion System	Throttle body	111
Power Train	Engine light	100

Table 4. 2-Gram Aggregated Result based on Model and Year

Model	Year	2-Gram	Count
Fusion	2010	Throttle body	327
Escape	2010	Throttle body	247
Escape	2008	Power steering	122
Fusion	2011	Throttle body	122
Escape	2011	Throttle body	117
Fusion	2010	Wrench light	108
Escape	2009	Throttle body	94
Escape	2008	Steering wheel	76
Escape	2010	Wrench light	76
Escape	2010	Gas pedal	66

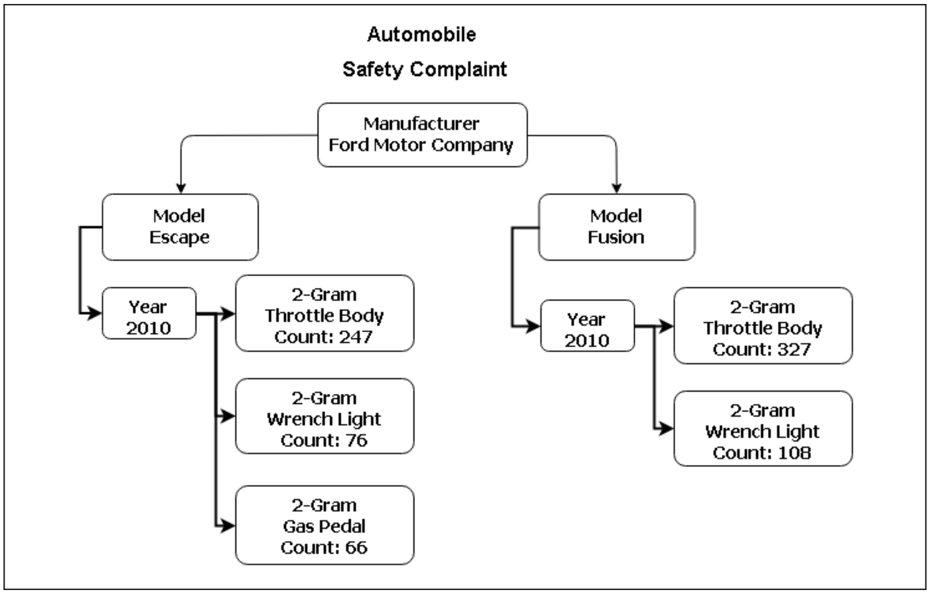

Fig. 1. Concept Map from Part of Table 4

CONCLUSION

This paper proposes a new approach of expanding the existing ontology by creating a new concept map for domain specific needs. The knowledge sources are from public available dataset generated by consumer on automobile safety. Our approach is implemented through SAS Text Miner 15.1 to extract detail level entities as new concepts and aggregate the result at different domain specific levels. These results are used to create concept map for further analysis.

REFERENCES

1. Novak, J. Canas, A.: The theory underlying concept maps and how to construct and use them. Technical Report. Institute for Human and Machine Cognition, Florida, 1-36 (2008).

2. Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R.: Ontologies for enterprise knowledge management. IEEE Intelligent Systems, 18(2), 26-33 (2003).

3. Cimiano, P. Völker, J.: text2onto. In: International conference on application of natural lan-guage to information systems, pp. 227-238. Springer, Berlin, Heidelberg (2005).

4. Vigo, M., Bail, S., Jay, C., Stevens, R.: Overcoming the pitfalls of ontology authoring: Strategies and implications for tool design. International Journal of Human-Computer Studies, 72, 835-845 (2014).

5. Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F. Cucchiarelli, R.: Extending and enriching WordNet with OntoLearn. In: Proceeding of 2nd Global WordNet Conf.(GWC), pp. 279-284. (2004).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sadikshya Basnet

Oakland University

248-8323431

Sadikshyabasnet21@gmail.com