https://doi.org/10.1109/ETFA.2019.8869057.

Optimized Automotive Fault-Diagnosis based on Knowledge Extraction from Web Resources

Simon Meckel*, Johannes Zenkert*, Christian Weber*, Roman Obermaisser*, Madjid Fathi*, Rubaiyat Sadat†
*University of Siegen, Germany
†Mulytic Labs, Munich, Germany

***Abstract*—The maintenance and repair of modern vehicles is a challenge for garages, as different causes of faults lead to similar symptoms in the highly complex vehicles these days. Existing processes for fault-diagnosis based on manufacturer service manuals and human experiences are often inadequate and result in high effort and wrong decisions. In addition to these service manuals which provide basic models for e.g., diagnostic terms, primary physical quantities, causal relationships, and plausibilities, nowadays, internet forums offer a comprehensive source of experiences for solutions to these challenges. This paper, therefore, presents methods for the extraction of knowledge from unstructured and informal contributions in internet forums with the goal to synthesize diagnostic graphs from the established knowledge base, which are part of a maintenance software to supports garages in the maintenance of vehicles by suggesting more efficient and target-oriented diagnostic and maintenance actions in real-time.**

***Keywords*—Automotive Fault-Diagnosis, Knowledge Extraction, Optimization**

I. Introduction

For garages, diagnosing faults in modern motor vehicles is a very time-consuming task, which in many cases takes more time than the actual repair. The increasing complexity of the vehicles and the integration of new technologies, services and components are leading to growing challenges of the maintenance in garages. On-Board Diagnostic (OBD) fault codes and customer fault symptom descriptions help technicians to identify and locate faults and select appropriate maintenance procedures. However, different faults may lead to similar symptoms and it is usually not immediately clear to which components a repair must refer. Consequently, fault diagnosis is necessary, i.e., narrowing down faults from first symptoms to their causes by tests and feedback from the tests.

The Association of British Insurers reports that the average cost of a car repair bill has risen by about one third over the last three years and refers this rise in costs mainly with new electronic systems and stricter security regulations. Vehicle diagnosis today is based on fault symptoms and service manuals from automobile manufacturers, which suggest various tests for workshop technicians to pinpoint the cause of a fault. However, this procedure has several disadvantages:

- Dependence of workshops on model-specific service manuals of automobile manufacturers.
- Complicated and lengthy test procedures that do not necessarily reveal the (correct) cause of the fault.
- Unsatisfactory percentage of fixed-first-visits with a problem solution during the first workshop visit, often requiring several visits until the fault cause is finally diagnosed and corrected.
- No direct feedback of service technician expertise and experience for better test procedures.
- Long downtimes due to lengthy diagnostic and repair processes are unpleasant for vehicle owners and the non-transparency of the procedure might lead to incomprehensible costs.

Internet forums offer significant potential beyond OEM service manuals for improved diagnostic procedures through the use of experience knowledge to provide a faster, more accurate and more reliable root cause fault analysis.

This paper proposes an innovative technique to harness the knowledge of laypersons, amateur technicians and professional authors sharing their diagnostic and repair experiences in relevant internet forums. In an automated manner, a knowledge base is formed which summarizes extracted knowledge in a structured form. Synthesis algorithms utilize this knowledge in combination with OEM repair strategies and customer symptom descriptions to synthesize a diagnostic graph from primary concepts such as error codes or sensor data, throughout the utilization of logical relationships to the end points, i.e., the identified root causes of faults. The paths between primary information to the identified root causes vary considerably with respect to the number and time of tests to be conducted. This leads to significant differences with regard to the effort for the technicians. Each created path is characterized by a certain confidence regarding the expected positive fault identification, such that this initial diagnostic graph needs to be optimized in a third step. Meaningful optimization parameters are the overall working time, the number of required technicians or the experience level of the technicians to perform certain tasks. As part of a maintenance software, the optimized diagnostic graphs support garage technicians in the maintenance of vehicles by suggesting more efficient and target-oriented diagnostic and maintenance actions in real-time.

II. Related Work

In order to achieve an improvement of diagnostic processes of vehicles in garages utilizing knowledge about vehicle diagnoses from internet forums, a cooperation of the two large specialist areas of technical vehicle diagnosis as well as text-related knowledge extraction is necessary.

978-1-7281-0303-7/19/$31.00 ©2019 IEEE

A. Fault-Diagnosis of Vehicles

The automotive industry has undergone a transformation from manual off-board testing to on-board diagnostics (OBD) [1]. Nevertheless, fault detection and diagnosis (FDD) techniques are largely concentrated on individual areas of the vehicle such as engine management, steering or braking and often do not take into account the interconnections of the components, i.e., the diagnosis is not performed on a system wide level. OBDs are integrated into the electronic control units (ECU) to indicate vehicle faults that are further evaluated off-board for troubleshooting. FDD approaches last from simple threshold detection up to learning algorithms for non-linear signal processing [2] and sensor fusion approaches [4]. Case based reasoning has been also investigated [8]. Recently, FDD techniques for the knowledge based detection of anomalies are explored [6]. All these techniques are conceptualized under the premise that OBD error codes are always linked to the faults of certain components, whereas practice shows that this dependency does by far not cover all available sources of information to track down occurred faults to their root cause. The current state of research lacks fully available sources for the fault descriptions and the integration of peer-learning-techniques into vehicle fault-diagnosis processes.

B. Knowledge Extraction and Integrative Text Mining

In [9] integrative text mining is proposed as a methodology for the combination and representation of different text analysis results, which are summarized in a multidimensional knowledge representation (MKR) for application in knowledge discovery and knowledge visualization. The CIMAWA algorithm is proposed in [7]. Its mathematical representation is

$$\mathit{CIMAWA}^{\zeta}_{ws}(x(y)) = \frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}} + \zeta \frac{Cooc_{ws}(x,y)}{(frequency(x))^{\alpha}}$$

The calculation method measures the word association strength between a term x and a term y, which occur within a text corpus as co-occurrence $Cooc_{ws}$ in a window size ws of 10 words. In the paper's knowledge extraction approach, we use the CIMAWA word association strength to determine related word pairs as described in Section III-B.

III. Knowledge Extraction

A. Webcrawling and Natural Language Processing

Figure 1 demonstrates the process of web crawling and natural language processing (NLP). Web crawling is used to transfer forum posts to a knowledge base. The texts in the relevant forum pages are extracted and stored together with metadata. The texts are then pre-processed by the NLP operations sentence splitting, part of speech tagging, and tokenizing. Using a lexical resource (here referred to as car thesaurus) these pre-processed texts are examined for relevant entities. Sentences containing the entity are then further explored using a part of speech tagger.

B. Extraction and Calculation of Word Associations

Components of the sentence, i.e., single words and especially verbs that indicate actions, are extracted and stored as co-occurrence pairs (Fig. 1, 3rd column). Using the CIMAWA algorithm and the corpus of all texts, the word association strengths between the entities and the surrounding words are determined. Finally, these entity process descriptions (e.g., *"unscrew cover", "unscrew screws", "measure stand", ...*) are extracted and stored as co-occurrence between entity and verb to describe the action.

In the next step, the source texts are again scanned for co-occurrences, so that an assignment of entity process description and source text exists. This is done for all pairs, so that in case of multiple occurrences of entity process descriptions chains are formed from these pairs. The order in which the pairs are named in the text is taken into account to ensure a logical sequence of the entity process descriptions by a large number of texts. A large number of texts containing similar entity process descriptions increases the confidence regarding the validity of the overall chain description.

By using phrases that describe sequences, possibilities, or alternatives, the entity process descriptions are arranged in a form that resembles an undirected graph, as shown in Fig. 2 (1st column). The descriptions are used as edges to move from one result, presented as a node, to another node. The entity process descriptions can be interpreted as tests. Every co-occurrence chain that is formed in this way has a starting point and an end point, which is the last result of a test.

On demand, additional meta information can be manually stored at the edges, i.e., at the entity process descriptions, after the extraction. By adding information (e.g., based on an experience data base) about test duration, number of necessary employees, difficulty of the process, required special equipment or costs, the edges get different weights or annotations that influence a possible selection of the subsequent tests. This results in an optimization problem with multiple objective functions.

IV. Diagnostic Graph Synthesis and Optimization

Based on the extracted co-occurrence chains the graph modeling algorithm detects phrases that indicate variants, alternatives or conditions and synthesizes the diagnostic graph, which is a directed acyclic graph. This is demonstrated in Fig. 2 (2nd column). The nodes of the graph, i.e., the tests, get reduced to explicitly workable tasks, e.g., measure a voltage at a certain location. The graph structures the tests as sequences to be performed by garage technicians, where their feedback after every test contributes to confidence values at the graph edges. The initial inputs of the diagnostic graph are given by the primary concepts of meaningful fault symptom descriptions such as OBD error codes or user symptom descriptions that match some extracted co-occurrence chains. For the graph synthesis, methods of graph and network theory as well as machine learning techniques [3] are applied. The extracted information from the internet forums provide additional key facts which allow to structure and rank the tests. Key phrases (e.g.,

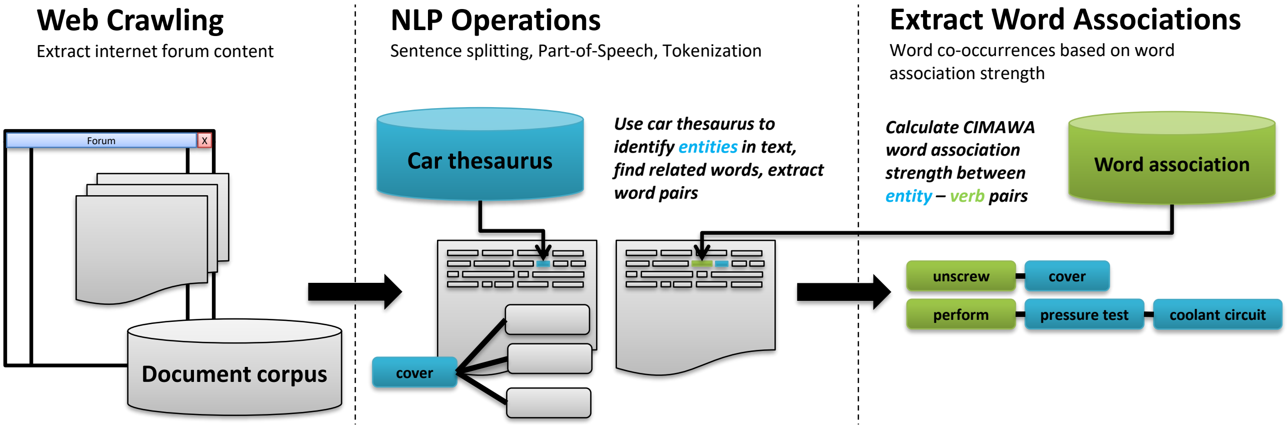
Web Crawling
Extract internet forum content
Forum
Document corpus
NLP Operations
Sentence splitting, Part-of-Speech, Tokenization
Car thesaurus
Use car thesaurus to identify entities in text, find related words, extract word pairs
cover
Extract Word Associations
Word co-occurrences based on word association strength
Calculate CIMAWA word association strength between entity – verb pairs
Word association
unscrew
cover
perform
pressure test
coolant circuit

Fig. 1. Process of web crawling and natural language processing

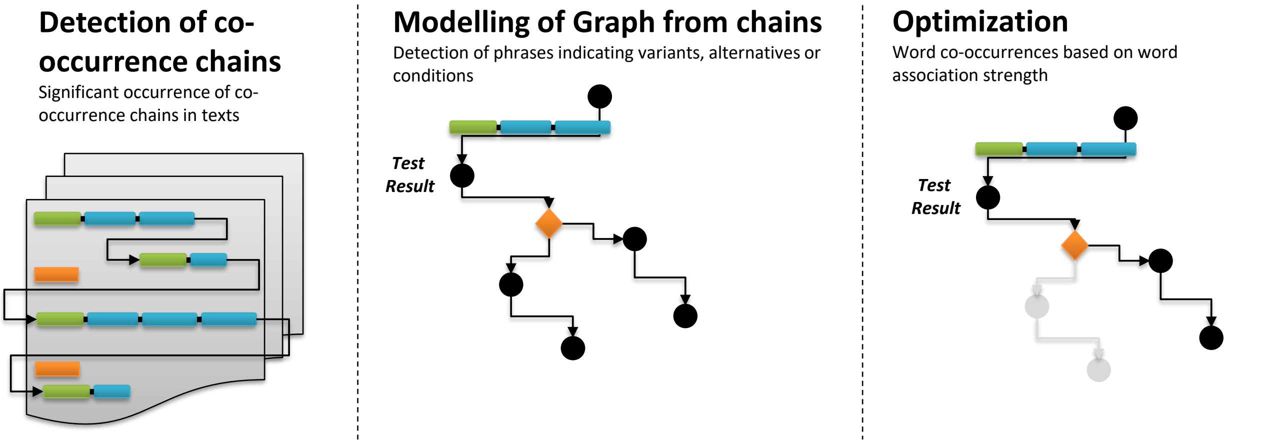
Detection of co-occurrence chains
Significant occurrence of co-occurrence chains in texts
Modelling of Graph from chains
Detection of phrases indicating variants, alternatives or conditions
Test Result
Optimization
Word co-occurrences based on word association strength
Test Result

Fig. 2. Process of the synthesis and subsequent optimization of the diagnostic graph

forum post marked as ”successful”) show a high confidence of a test sequence. Furthermore, frequently mentioned tests (or test sequences) as well as often referenced tests typically offer a higher reliability, than tests where users’ feedback attest a lower success rate. The graph synthesis neglects information regarding repetitions of tests or dispensable tests such that the resulting diagnostic graph offers multiple paths from the starting nodes to the end nodes, i.e., the potential root causes of the fault (Fig. 2, 2nd column). The last step to establish the final diagnostic test structure for garage technicians is an optimization of the multi-path graph after the synthesis. For this, the optimization algorithm evaluates key phrases and the assigned attributes, and classifies the tests according to their precursors and successors, in this way being able to spot redundant tests and finding shortcuts in the overall graph (Fig. 3b). The measure of the modularity [5] of a graph shows, how good it can be split into stable sub-graphs. Depending on the desired target function, e.g., shortest time optimization or fewest workers number, it calculates an optimized solution in form of a sequence of tests to be conducted for a sound root cause analysis.

V. Example Use Cases

In this section we exemplarily demonstrate the working principle of our proposed algorithms by means of a realistic scenario where a first fault symptom of a vehicle indicated by an OBD error code leads to a visit of a garage, where the root cause of the fault is further analyzed. The use case shows the significant improvement of structured test sequences over an experience-based strategy. In the current state of our project several steps are conducted manually, e.g., some assignments of attributes and graph optimizations, however, this does not affect the general working principles.

Fault Symptom from Engine Coolant Level Sensor

In our scenario we assume that the engine coolant level warning indicator is activated. To solve such a problem, nowadays, a typical auto repair shop tries out different tests or repairs, often in an unsystematic manner, where another workshop with its own experts may proceed differently.
In contrast to this, our MKR, extracted from an internet forum, allows to generate a diagnostic graph that matches the first symptom as the graph’s starting point.

Figure 3a shows a potential outcome of the diagnostic graph synthesis before the optimization and Fig. 3b shows an

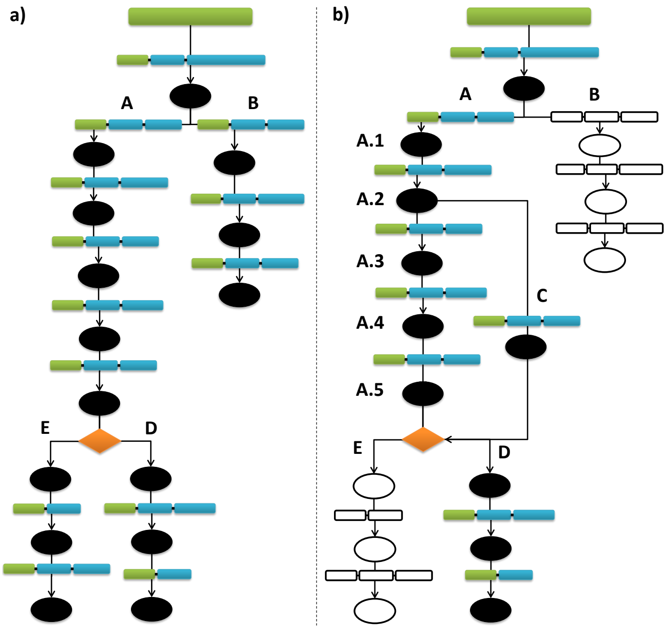

Fig. 3. Example of a synthesized and optimized diagnostic graph

optimized version. The difference between the two versions of the graph is that the first one simply lists useful situation-related tests for the identification of the root cause of the fault, resulting in a longer and more cost intensive diagnosis procedure. The optimized version restructures, summarizes, and disregards certain tests by evaluating all assigned attributes (confidence, time, difficulty, ...) and produces the solution according to the desired target function, where we choose time efficiency. For example, the necessary tests include leaking tests for many involved components like the hoses, tank, clamps, or seals. Based on the confidence values and constraints, all these tests can be more beneficially performed if the cooling circuit is put under pressure, consequently suggesting the pressure test as the second test after a first general visual inspection of the liquid level. We expect good service technicians to also have this knowledge, however, our algorithms provide a multi-source synthesis of available information, which rates the quality of the diagnostic graph and which scales when the pool of knowledge becomes bigger.

The outcome of each test can include or exclude certain root causes, e.g., if the coolant level is correct despite the OBD indication, the error code is wrong. The pressure test consequently summarizes the inspection of the tightness of all accessible components (path C). The outcome of a test is often dependent on the situation in which it is performed, e.g., close to the components' operating temperatures or at room temperature, and thus, the tightness test might only reveal a problem with a seal if the test conditions are fitting. At path D in Fig. 3b, the most probable steps that the algorithm proposes are an in-depth inspection of the tightness of the high pressure pump during operation as well as under idle conditions. The other path, E, was rated less likely and might still lead to a correct solution if path D turns out to be unsuccessful.

VI. Conclusion and Future Work

In this paper we motivate the usage of knowledge extraction and integrative text mining in combination with diagnostic graph synthesis and optimization in order to exploit the comprehensive knowledge of internet forums. This solution goes beyond the OEM service manual instructions, to improve the diagnostic procedures for car repair garages. Our approach consequently makes the accumulated knowledge of a broad range of experienced contributors available for garage technicians such that a structured, transparent, and thus, standardized diagnostic test procedure becomes possible. With an example we show that in the current state of our project it is possible to create a multidimensional knowledge representation from relevant internet forums and that our graph synthesis is able to automatically derive optimized diagnostic strategies.

The next steps of our project include the extension of the MKR and fully automated algorithms for the diagnostic graph synthesis. Future research aspects of the project will especially concentrate on generic diagnostic models and model transformations with the goal to include car model-specific differences into our diagnostic graphs, eventually being able to merge core elements (valid for all vehicles) with brand and model-specific graph elements for an up-to-date fast and reliable vehicle diagnostic procedure.

We see another use case for the proposed approach in the area of industrial automation. High potential lies in the analysis of textual process documentation [10], which can also be transformed into diagnostic graph structures.

References

[1] P. Engelke and H. Obermeir. Funding project diana—integrated diagnostics for the analysis of electronic failures in vehicles. In *Test Symposium (ETS), 2012 17th IEEE European*, pages 1–1. IEEE, 2012.

[2] J. P. N. González. Vehicle fault detection and diagnosis combining an aann and multiclass svm. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 12(1):273–279, 2018.

[3] S. Meckel, R. Obermaisser, and J.-U. Yang. Generation of a diagnosis model for hybrid-electric vehicles using machine learning. In *2018 21st Euromicro Conference on Digital System Design (DSD)*, pages 389–396. IEEE, 2018.

[4] S. E. Muldoon, M. Kowalczyk, and J. Shen. Vehicle fault diagnostics using a sensor fusion approach. In *SENSORS, 2002 IEEE*, volume 2, pages 1591–1596 vol.2, June 2002.

[5] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[6] A. Theissler. Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. *Knowledge-Based Systems*, 123:163–173, 2017.

[7] P. Uhr, A. Klahold, and M. Fathi. Imitation of the human ability of word association. *International Journal of Soft Computing and Software Engineering (JSCSE)*, 3(3):248–254, 2013.

[8] Z. Wen, J. Crossman, J. Cardillo, and Y. L. Murphey. Case-base reasoning in vehicle fault diagnostics. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 4, pages 2679–2684 vol.4, July 2003.

[9] J. Zenkert, A. Klahold, and M. Fathi. Knowledge discovery in multidimensional knowledge representation framework. *Iran Journal of Computer Science*, 1(4):199–216, 2018.

[10] J. Zenkert, C. Weber, A. Klahold, M. Fathi, and K. Hahn. Knowledge-based production documentation analysis: An integrated text mining architecture. In *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 717–720, Aug 2018.