Artificial Intelligence for Engineering Design, Analysis and Manufacturing (2017), *31*, 313–326.
© Cambridge University Press 2016 0890-0604/16
doi:10.1017/S0890060416000470

Learning to predict characteristics for engineering service projects

LEI SHI,[1] LINDA NEWNES,[1] STEVE CULLEY,[1] AND BRUCE ALLEN[2]

[1]Department of Mechanical Engineering, University of Bath, Bath, United Kingdom
[2]AIRBUS Operations Ltd, Filton, United Kingdom

(RECEIVED October 1, 2015; ACCEPTED May 31, 2016)

Abstract

An engineering service project can be highly interactive, collaborative, and distributed. The implementation of such projects needs to generate, utilize, and share large amounts of data and heterogeneous digital objects. The information overload prevents the effective reuse of project data and knowledge, and makes the understanding of project characteristics difficult. Toward solving these issues, this paper emphasized the using of data mining and machine learning techniques to improve the project characteristic understanding process. The work presented in this paper proposed an automatic model and some analytical approaches for learning and predicting the characteristics of engineering service projects. To evaluate the model and demonstrate its functionalities, an industrial data set from the aerospace sector is considered as a the case study. This work shows that the proposed model could enable the project members to gain comprehensive understanding of project characteristics from a multidimensional perspective, and it has the potential to support them in implementing evidence-based design and decision making.

Keywords: Characteristics Learning; Characteristics Prediction; Engineering Service Project; Knowledge Discovery

1. INTRODUCTION

Many service providers are now offering maintenance and repair of long-life, high-value manufacturing products, often the original equipment manufacturer such as Boeing, Airbus, and Rolls-Royce (Baines et al., 2007). An example of such engineering services can be found in maintenance repair and overhaul (MRO) service providers for airline fleets. According to the International Air Transport Association, the global market for such providers was estimated at $50 billion in 2011, an increase of 11% compared with 2010 (http://www.iata.org). MRO service providers often analyze trends, features, and failure for products, to enhance their operations and effectiveness. However, one of the challenges faced is that the service occurs globally, consists of complex service systems, often has distributed teams/experts, and is supported by data that is heterogeneous by nature (Zhen et al., 2011).

The aim of the research described in this paper is to demonstrate how companies such as MRO providers can enhance their decision making and deliver more timely and cost effective services through the automatic analysis of engineering project data.

Within this paper a review of the current state of the art in data analytics is provided and shows that there is a need to provide tools and techniques to automatically analyze services entering an MRO facility. In order to explore the potential of using historical data and modeled knowledge to support the understanding of project characteristics, this paper proposes a characteristic learning and prediction model. The model incorporates with some analytical approaches by using data mining and machine learning techniques. It shows how projects can be characterized, such as level of activity, level of complexity, and sequence of activities.

Through the use of the analytical approaches, the model aims to provide automatic project characteristic identification, representation, and prediction, and support project members in implementing evidence-based design and decision making. To evaluate the approaches and demonstrate their functionalities, this paper describes the approach through the use of an industrial data set from the aerospace maintenance repair and overhaul sector. The data set focuses on unplanned repairs for airline fleets.

The paper is organized as follows: Section 2 reviews related work. Section 3 introduces the characteristic learning and prediction model and related analytical approaches. Section 4 describes the industrial data set and demonstrates the use of the model as a learning and prediction tool. Section

Reprint requests to: Lei Shi, Department of Mechanical Engineering, University of Bath, Bath BA2 7AY, UK. E-mail: l.shi@bath.ac.uk

5 concludes the paper and discusses the future activities of the research.

2. RELATED WORK

Often large data sets are difficult for individuals to assess, and information overload often prevents the effective reuse of project data and knowledge (Paroutis & Al Saleh, 2009). This can make the understanding of project characteristics difficult (Griffin, 1997).

Many engineering service projects are highly collaborative, interactive, and distributed. Hence, project members would be unable to review all the data being captured through the product life cycle, due to the restriction of time and ability. For most of them, gaining a comprehensive understanding of project characteristics could be a challenging task. This is evidenced by the research being undertaken in domains such as product data management (Mesihovic et al., 2004; Feng et al., 2009).

The service design of engineering products utilizes various types of data and information such as the product features, functionalities, and reliability, as well as the customer requirements, market demands, and the organizational context of the service (Petersen et al., 2003; Luchs & Swan, 2011).

In the design of engineering services, a thorough understanding of the characteristics of previously completed service projects is important (Shi, Gopsill, Snider, et al., 2014). These characteristics cover the aspects of engineering process, project performance, activity sequence, communication content, the scale of collaboration, and resource consumption. Understanding these multiple characteristics and their inner relationships is a complex process, but one which is considered critical for the design engineers to perform the process design, task planning, resource allocation, and decision-making related work in an effective manner (Shi, Gopsill, Newnes, et al., 2014; Snider et al., 2014).

Numerous studies have shown that the data and knowledge of historical service projects can be used as guidance for the design of future engineering services (Baines et al., 2007; Doultsinou et al., 2009; Settanni et al., 2015). The work by Zhang et al. (2012) describes an engineering service approach for large construction machines. Their findings describe how it is important to analyze the design, in-use, and failure data for the product and the supporting of integrated services. This is also required when designing more complex engineering products and services such as an aircraft. Hence, the service knowledge and generic service process regarding product components are necessary to be identified from the historical service data and records (Shi et al., 2015).

Research has also demonstrated that these complex engineering service systems integrate people, processes, and products (Chuang, 2007; Goh et al., 2015). Because large-scale engineering projects include multiple types of characteristics that cross disciplines and functions, their data sets are often distributed and interpreted by the project members within different areas of expertise (Li et al., 2009). To achieve the effective administration and management for such projects, the product and process knowledge may need to be captured from the communication data, individual behaviors, and team interactions (Wasiak et al., 2011). To this end, it is important that the understanding of project characteristics should consider multiple perspectives.

2.1. Understanding engineering project characteristics

The understanding of project characteristics has a direct influence on project success. The comprehensive understanding could help project members optimize the project structure, refine the project granularity settings, and make rational decisions on product design, manufacturing, and marketing (Cho et al., 2009; Li, Xie, et al., 2009).

In practice, project characteristics often have various definitions according to information needs. From an engineering design perspective, they are defined as the high-level interpretations of project features, which are associated with the project operation and management tasks (Snider et al., 2014). For certain types of projects, their characteristics can be defined in explicit forms. The characteristics of software engineering projects include, for example, the design paradigm, programming language, database type, and amount of created source code (Walter, 2014). However, project characteristics can also be defined in implicit forms. For example, from a management perspective, project characteristics usually contain performance indicators such as elapsed time and cost (Cho et al., 2009). Meanwhile, the uncertainty and complexity are also considered as the key aspects of project characteristics (Ahmad et al., 2013). From an information management perspective, the trace of digital object creation and evolution can be considered a characteristic enabling project members to have a detailed understanding of engineering design processes (Gopsill et al., 2014). From a human–computer interaction perspective, the change of sentiment or topic of communications provides project characteristics aimed at assisting project members to monitor work complexity and track project progress (Jones et al., 2015).

Engineering projects often involve multiple collaborators, utilize distributed resources, and handle heterogeneous and fragmented data. Consequently, their characteristics can be dynamic and time sensitive in nature. Understanding the evolving characteristics can be challenging for most project members (Engwall & Jerbrant, 2003; Meredith & Mantel, 2011). In most cases, the processing of project data, capturing, and modeling the project knowledge is often undertaken using expert effort to assess, evaluate, understand, and interpret the information into a usable form (Pascal et al., 2013). This understanding and identification of characteristic is largely dependent upon the availability, capability, and experience of the project members. The following issues can occur due to these dependencies:

- *High cost on knowledge discovery and capture:* In the characteristic understanding process, the knowledge concepts contained by the project data need to be discovered and captured. To perform this task, data items created during the project operation should be reviewed and analyzed. The identified knowledge concepts are often organized through the use of knowledge models, such as ontology, concept map, or semantic web. Relying on manual work means this task requires intensive efforts from the knowledge experts or experienced project members, implying it can be considerably time consuming and expensive.

- *Diversified perspectives of characteristic understanding:* In order to understand the project characteristics at the appropriate level of granularity, project members are required to have the knowledge about the project components and the interrelationships between them (Shi, Gopsill, Snider, et al., 2014). Such knowledge could be the specific and precise descriptions of tasks, subsystems, project teams, the sequence of multiple tasks, the dependency of multiple subsystems, and the responsibilities of different project teams. Within a service project, numerous project members and project components are involved. Chandrasegaran et al. (2013) describe how knowledge sharing and utilzation among the project teams can be difficult. Various knowledge capabilities of the project members could lead to different viewpoints/understanding for identical project components (Chungoora et al., 2013).
- *Low capability of processing complex data:* A service project could generate various types of data, for example, e-mails, instant messages, formal reports, spreadsheets, computer-aided design models, and simulation files. During the characteristic learning process, the project members need to spend time on reviewing and analyzing such data. However, due to the restrictions on time and capability, processing large amount of complex data using manual work is not always realistic. More important, such data is likely to be generated by multiple sources, and have heterogeneous and fragmented forms, so that the manual data processing can cause the analysis results to lack accuracy or be problematic. This can be exacerbated if the project members lack certain knowledge, or when human errors occur.

To overcome these challenges, the analytical techniques such as data mining and machine learning need to be adopted.

2.2. Data mining and machine learning based models

With the application of information and communication technology, large volumes of data can be generated, captured, and stored during the project execution process. The data contains detailed information about project objectives, processes, outcomes, problems encountered, and lessons learned, so that it can be used to facilitate the understanding of project characteristics (Harding et al., 2006; Choudhary et al., 2009).

To process the data in an effective manner, the use of data mining and machine learning techniques is considered to be critical. Data mining includes specifically designed computational approaches that could achieve automatic knowledge discovery, condition monitoring, pattern recognition, and association analysis. Meanwhile, machine learning includes advanced statistical methods that could achieve automatic decision making, predictive modeling, data classification, and data clustering. In practice, both techniques have been applied in various fields to create intelligent systems or analytical approaches (Chen et al., 2012; Kamsu-Foguem et al., 2013).

As stated by recent research, there is an increasing trend of using data mining and machine learning in manufacturing fields (Wang, 2007; Köksal et al., 2011; Wu et al., 2012). Numerous applications and models have been proposed and introduced by the research community, with the aim of achieving engineering knowledge management, production process management, project characteristic identification, and performance monitoring.

To capture the process-related knowledge and identify the sequential activities of engineering projects, Shi, Gopsill, Newnes, et al. (2014) proposed an analytical approach based on data mining and nature language processing techniques. The approach extracts knowledge concepts from engineering documents, and automatically interprets activity sequences of engineering projects. The output of this approach provides pictorial and numeric knowledge representations to the project members. The aim is to enable the project members to learn the structures of different projects processes, and also to help them understand the similarity and normality of such processes.

To understand the characteristics of the engineering design process, Gopsill et al. (2014) proposed an analytical tool. It applies metadata analysis and frequency analysis to identify the evolution of product design activities, that is, the creation and modification of computer-aided design and computational fluid dynamics files. By using this approach, the project members are able to assess the quantity of generated digital objects at different project stages, and to understand the dependency changes of such files over time.

For MRO activities, the analysis of service records would assess the product condition and predict the fault. Certain machine learning techniques such as support vector machine, artificial neural networks, and decision trees are considered useful. These data classification and decision-making approaches can be used to validate the product running status on a real-time basis by analyzing the operating data and historical data. Once any unexpected status of the product has been detected, these approaches can make intelligent decisions in an automatic manner (Widodo & Yang, 2007; Kankar et al., 2011).

To design the preventive maintenance solutions for complex engineering products or infrastructures, the integration of data mining and machine learning is necessary. For instance, to diagnose the fault of aircraft engines, Sahin et al. (2007) proposed a model that integrates Bayesian networks,

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

particle swarm optimization, and parallel computing. Moreover, to predict the risk of failures of a power grid, Rudin et al. (2012) introduced a system framework that employs supervised learning, ranking, and mean time between failures modeling.

From the examples described in this section, there are various information needs that are dependent upon an individual's role; hence, the creation of a model for learning and predicting the characteristics of collaborative and complex engineering projects may need to apply the combinations of techniques. However, presently there is still limited research describing how to select appropriate techniques to create the model and assist in the analysis of heterogeneous engineering data (Wagstaff, 2012). Hence, the adaptive capability and reusability of such models still need to be improved.

3. PROPOSED MODEL AND APPROACHES

The core modules of the proposed model are characteristic learning and characteristic prediction. As shown in Figure 1, the learning module identifies pattern-related, process-related and knowledge-related characteristics from historical project data. It categorizes such characteristics, and then sends them to the prediction module. The prediction module treats the categorized characteristics as the training data, and then applies them to predict the characteristics of ongoing projects. During the learning and prediction processes, a feature-based knowledge representation of each project is created. The creation of such knowledge representation involves knowledge-based feature selection and characteristic representation modules. The detailed information of this model and involved approaches is given in the following sections.

3.1. Feature modeling

During the characteristic learning process, a project is represented using a multilayer structure (see Fig. 2). From top to bottom, the layers contain project layer, characteristic layer, feature layer, and data layer. For any information elements contained in the adjacent layers, a one-to-many relationship is used to connect them. Based on this structure, a project can be represented by a set of hierarchical characteristics: each characteristic is represented by a set of features, and each feature is derived from the content of multiple data items.

In the data layer, the data items are gathered from different project teams. To enable the application of data mining and machine learning techniques, all these data items should be stored in digital formats.

In the feature layer, all the project features are extracted by analyzing the content of data items. Project feature is the atomic information element in this model. Each project feature contains a meaningful description, and such description will be used to distinguish one feature from the others, or link the feature to the others. The fact that the feature does not have

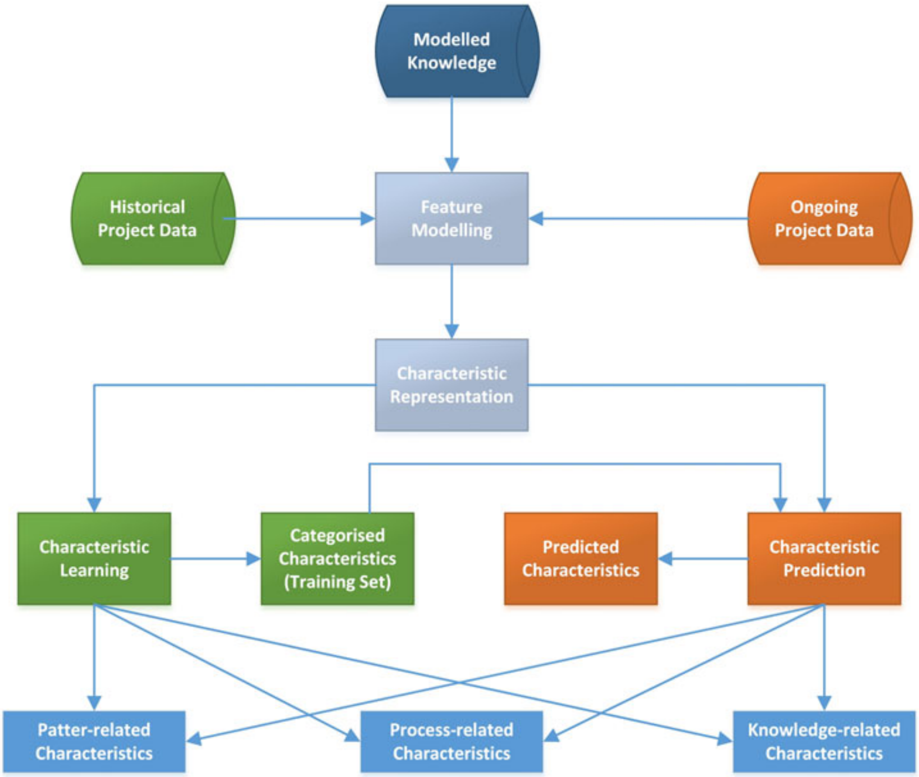

Modelled Knowledge
Historical Project Data
Feature Modelling
Ongoing Project Data
Characteristic Representation
Characteristic Learning
Categorised Characteristics (Training Set)
Predicted Characteristics
Characteristic Prediction
Patter-related Characteristics
Process-related Characteristics
Knowledge-related Characteristics

Fig. 1. The modules of the model.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

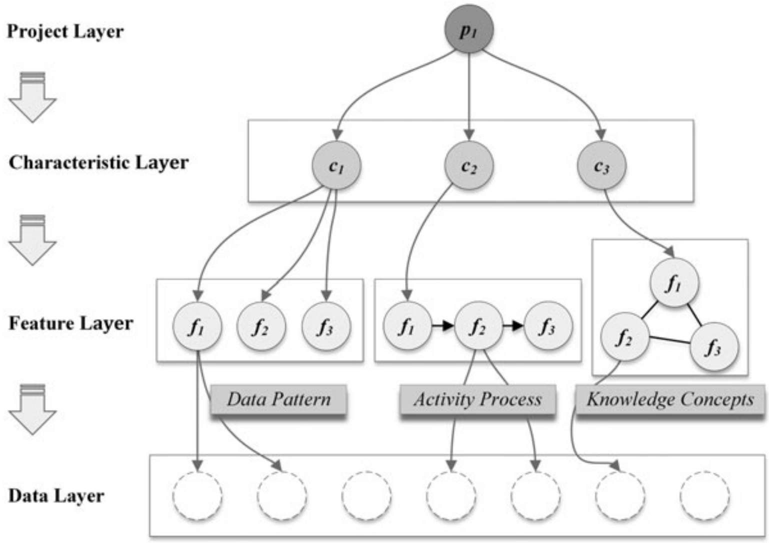


Fig. 2. A multilayer project representation.

format restrictions means it can be represented by a term, phrase, image, or file.

In the characteristic layer, a project characteristic comprises a set of features with interrelationships. These relationships are identified by analyzing the feature attributes and contexts. In this work, project characteristics are defined to have three forms: pattern related, process related, and knowledge related. The features contained in a pattern-related characteristic are independent of each other. For example, *the participation of team members* is a pattern-related characteristic; *the list of equipment names* is also a pattern-related characteristic. Such features are identified by counting the occurrence of people/equipment names contained in certain data items. The features of a process-related characteristic are time dependent. For example, *the fatigue test activity 1* and *the fatigue test activity 2* comprise a process-related characteristic. Such features are the adjacent activities being performed during the project. The features contained in a knowledge-related characteristic are time dependent, and they also have semantic associations with each other. For example, *the assembly activity 1, the assembly activity 2, the constrain of assembly activity 1*, and *the constrain of assembly activity 2* comprise a knowledge-related characteristic. Such features include the activities, the associations of the activities, and the constrains of the activity implementation.

Feature modeling is an automatic approach for extracting features from the project data. It contains three subtasks:

- *Feature identification:* It identifies potential features from the data content, and then organizes them as a feature set.
- *Feature selection:* It selects certain features from the feature set, and eliminates the ones with low significance.
- *Feature matching:* It identifies the relations among the selected features, and then organizes them as feature tuples.

The feature identification task applies named entity recognition (NER) and natural language processing (NLP). NER is used to automatically identify the named entities from the project data. Named entities indicate the terms or phrases related to date, location, organization, name of people, and name of artifact. NLP contains multiple subtasks, that is, tokenization, part of speech tagging, stop words removal, and stemming. As a fundamental task of content analysis, NLP is used to covert textual information into meaningful tokens or phrases.[1]

To understand the feature identification process, a sample data item is shown in Figure 3. It is a part of the report being generated by an aircraft service project. After implementing the feature identification, certain named entities and tokens are identified from the data content (see Table 1). Three types of operations, that is, regular expression, NER, and NLP, are applied. For example, the time-related information elements with international standard date format YYYY-MM-DD can be identified by using the following regular expression:

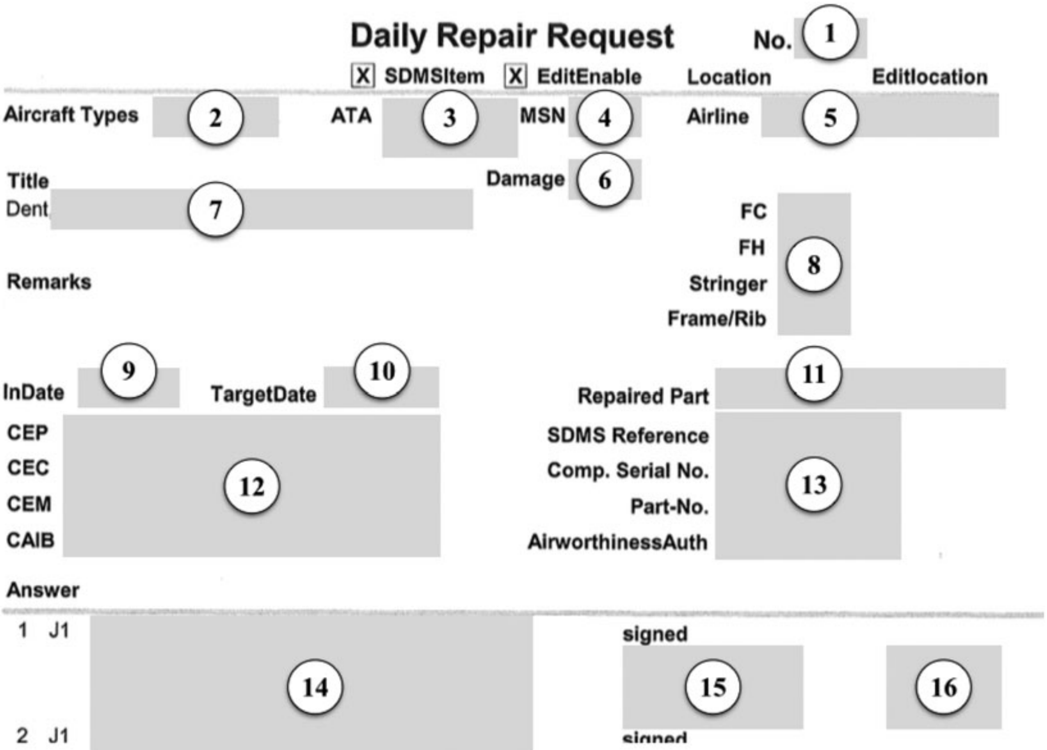
$$\begin{aligned}&\text{“}((?:(?:[1]\{1\}\backslash\backslash d\{1\}\backslash\backslash d\{1\}\backslash\backslash d\{1\})\backslash(?:[2]\{1\}\backslash\backslash d\{3\})))\\&\quad\times[-:\backslash\backslash/.](?:[0]?[1-9]\backslash[1][012])[-:\backslash\backslash/.]\\&\quad\times(?:(?:[0-2]?\backslash\backslash d\{1\})\backslash(?:[3][01]\{1\})))(?![\backslash\backslash d])\text{”}\end{aligned}$$

The artifact name *fuselage* and the people name *Bob Lewis* can be identified by using NER. The identification methods regarding different types of features are shown in Table 1.

After this process, the identified elements are treated as the potential features of the project, which will be processed by the feature selection task in the next step. In this model, only some of the identified named entities or tokens will be

[1] In this work, the open-source toolkit NLTK is applied to implement the natural language processing related tasks. It can be downloaded from http://www.nltk.org.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press



Daily Repair Request No.

☒ SDMSItem ☒ EditEnable Location Editlocation

Aircraft Types ATA MSN Airline

Title Damage

Dent

FC

FH

Remarks Stringer

Frame/Rib

InDate TargetDate Repaired Part

CEP SDMS Reference

CEC Comp. Serial No.

CEM Part-No.

CAIB AirworthinessAuth

Answer

1 J1 signed

2 J1 signed

Fig. 3. A data item generated by an aircraft service project.

recognized as project features. Feature selection is a variable selection method that is applied to select the essential variables from a given data set based on predefined rules or modeled knowledge (Gheyas & Smith, 2010; Shi & Setchi, 2013). It could filter out the features with low significance, so that the model could avoid the overfitting problem and deal with lower dimensional data.

In practice, the implementation of service projects often requires domain-specific knowledge. To identify the significance of features, the feature selection task should also apply the same knowledge. Such knowledge also helps the feature matching task detect the relations among the features. After this process, the selected features will be organized as feature tuples. For each project, a set of feature tuples will be then generated.

3.2. Feature-based characteristic representations

In this work, feature tuples from different projects are required to be interpreted using normalized data representations, for example, vector-based representation and sequence-based representation. Let p denote a project, f denote a project feature,

Table 1. *Features with descriptions*

Index	Description	ID Method	Index	Description	ID Method
1	Project ID	Regular expression	9	Creation date	Regular expression
2	Product model	NER Regular expression	10	Target date	Regular expression
3	Universal location number	Regular expression	11	Part name	NER
4	Manufacturer serial number	Regular expression	12	Manager name	NER
5	Customer information	NER NLP	13	Detailed part information	NER, NLP
6	Service type	NER	14	Detailed solution	NER NLP
7	Service description	NER NLP	15	Engineer name	NER
8	Product operation time	Regular expression	16	Completed date	Regular expression

Note: NER, Named entity recognition; NLP, natural language processing.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

τ denote a feature tuple, and K denote a knowledge base. To represent pattern-related feature tuples, the bag-of-words model and vector space model are applied,

$$\tau_{\text{pattern}} = [w(f_1), w(f_2), \ldots, w(f_n)]', \tag{1}$$

where w is a function to normalize the feature frequency.

To represent process-related feature tuples, the sequence-based data representation is applied,

$$\tau_{\text{process}} = g(f_{1,t_1}, a_1), \ldots, g(f_{n,t_n}, a_n), \tag{2}$$

where t_n is the timestamp of the feature, and $t_n > t_{n-1}$; g is a function to map the feature f_{n,t_n} to a specific activity a_n.

To represent knowledge-related feature tuples, the knowledge-based feature vector is used,

$$\tau_{\text{knowledge}} = [\varphi(f_{1,t_1}, K_1), \ldots, \varphi(f_{n,t_n}, K_n)]', \tag{3}$$

where φ is a function to map the feature f_{n,t_n} to a set of knowledge concepts K_n, where $K_n \subseteq K$.

A project characteristic could include multiple feature tuples. These tuples should explicitly describe the meaning of this characteristic. Let c_N denote a project characteristic with a specific meaning, for example, task complexity level; it could be represented by a collection of feature tuples,

$$c_N = \{(\tau_1, \ell_N), \ldots, (\tau_k, \ell_N)\}, \tag{4}$$

where ℓ_N is the label that indicates the meaning of c_N.

Based on the multilayer structure mentioned previously, a project can be represented by a combination of characteristics; that is, $p = \{C_{\text{pattern}}, C_{\text{process}}, C_{\text{knowledge}}\}$, where C_{pattern}, C_{process}, and $C_{\text{knowledge}}$ are the sets of different characteristics.

3.3. Characteristic learning process

To understand the meaning of feature tuples, the characteristic learning approach applies two strategies: labeling the unlabeled feature tuples based on prior knowledge, for example, based on a set of prelabeled feature-tuples; and labeling the feature tuples based their similarities (it is for the situation that the prior knowledge is insufficient).

With the first learning strategy, let c_N denote a project characteristic with label ℓ_N; c_N contains a set of labeled tuples, that is, $c_N = \{(\tau_1, \ell_N), \ldots, (\tau_k, \ell_N)\}$. For any unlabeled feature tuple (τ_i, ℓ_0), the label ℓ_N will be assigned to τ_i, if τ_i could satisfy $\tau_i \in c_N$, or τ_i and τ_j satisfy $\text{sim}(\tau_i, \tau_j) > \alpha$, where $\tau_j \in c_N$, α is a defined threshold. In other words, if an unlabeled feature tuple is included by an identified project characteristic, then the label of the characteristic will be assigned to this tuple. If the feature tuple is not included by any identified project characteristic, but it is similar to some tuples of that characteristic, then the label of the characteristic will be assigned to this tuple.

The second learning strategy is a semisupervised approach that means it could work with insufficient prior knowledge. Given a set of unlabeled feature tuples $X = \{(\tau_1, \ell_0), \ldots, (\tau_n, \ell_0)\}$, using clustering approaches, these tuples will be categorized into j clusters, that is, $X = \{C_1, \ldots, C_j\}$, where $C_j = \{(\tau_m, \ell_0), \ldots, (\tau_n, \ell_0)\}$. The feature tuples contained by the same cluster often have high similarity with each other in nature; thus, all these tuples should represent an identical project characteristic. The learning process then only needs to assign a label to the cluster, that is, $\ell_N \rightarrow C_j$, instead of each feature tuple.

To implement the learning strategies, similarity measure approaches are required. For the vector-based representation, the similarity measure is based on,

$$\text{sim}(\tau_i, \tau_j) = \frac{\tau_i \times \tau_j}{||\tau_i||\,||\tau_j||}. \tag{5}$$

For the sequence-based representation, an edit-distance based approach is applied (Shi, Gopsill, Newnes, et al., 2014),

$$d(\tau_i, \tau_j) = \min\begin{cases} \vartheta(\tau_i, \tau_j) + d(\tau_{i-1}, \tau_{j-1}) \\ \vartheta(\tau_i, \varepsilon) + d(\tau_{i-1}, \tau_j) \\ \vartheta(\varepsilon, \tau_j) + d(\tau_i, \tau_{j-1}) \end{cases}, \tag{6}$$

where ϑ is a cost function; $d(\varepsilon, \varepsilon) = 0$; and $d(\tau_i, \tau_j) = 0$, if τ_i and τ_j are identical.

The sequence similarity is calculated by using,

$$\text{sim}(\tau_i, \tau_j) = \begin{cases} 1 - \dfrac{d(\tau_i, \tau_j)}{\min(|\tau_i|, |\tau_j|)}, & \text{if } \alpha \le d(\tau_i, \tau_j) \le \beta \times \min(|\tau_i|, |\tau_j|) \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

where $|\tau_i|$ and $|\tau_j|$ indicate the length of τ_i and τ_j; α and β are defined thresholds, where $0 \le \alpha \le \beta \times \min(|\tau_i|, |\tau_j|)$, $\beta \in [0, 1]$. For example, when β is 0.5, the similarity of τ_i and τ_j will be considered as 0, if the edit distance between them is greater than the half-length of the shorter one.

3.4. Characteristic prediction process

In practice, the data generation of a project is a gradual process. It implies that the prediction of characteristics for an ongoing project at its early stage needs to utilize incomplete data.

Under this circumstance, the training data applied by the characteristic prediction process should reflect the dynamic relations among the time, data, and evolution of project characteristics. Assume a project p is contained by the training data, which has a set of characteristics, that is, $p = \{c_1, \ldots, c_N\}$, and each of the characteristics contains a set of feature tuples, that is, $c_N = \{(\tau_1, \ell_N), \ldots, (\tau_k, \ell_N)\}$. To perform the prediction on a real-time basis, the representation of characteristic needs to be segmented based on predefined time intervals. Consequently, a single characteristic will have multiple representations, each of which links to a specific project

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

stage. For a characteristic c_N, its segmented representations are represented as,

$$
\begin{aligned}
c_{N,t_1} &= \{\overbrace{(\tau_1, \ell_N), \ldots, (\tau_n, \ell_N)}^{t_1}\},\\
c_{N,t_2} &= \{\overbrace{(\tau_1, \ell_N), \ldots, (\tau_n, \ell_N)}^{t_1}, \overbrace{(\tau_1, \ell_N), \ldots, (\tau_n, \ell_N)}^{t_1+\Delta t}\},\\
&\ldots\\
c_{N,t_k} &= \{\overbrace{(\tau_1, \ell_N), \ldots, (\tau_n, \ell_N)}^{t_1}, \ldots, \overbrace{(\tau_1, \ell_N), \ldots, (\tau_n, \ell_N)}^{t_1+(k-1)\Delta t}\},
\end{aligned}
\tag{8}
$$

where Δt is the length of time interval.

For a project, it is represented by using the time-segmented characteristics,

$$
\begin{aligned}
p_t = \{&\{c_1, \ldots, c_N\}_{t_1}, \{c_1, \ldots, c_N\}_{t_1+\Delta t}, \ldots,\\
&\{c_1, \ldots, c_N\}_{t_1+(k-1)\Delta t}\},
\end{aligned}
\tag{9}
$$

where k is the number of time intervals.

As an example, Figure 4a shows the process of a service project. This project has three different stages: planning stage, problem-solving stage and evaluation stage. The data items regarding each stage have various file format and storage locations. In this case, each stage takes approximately 30% of the project progress. The process segmentation therefore is implemented based on the 30% interval; that is, the project process will be segmented into three subprocesses that map to the stages accordingly. As shown in Figure 4b, each subprocess is used to represent the project characteristics.

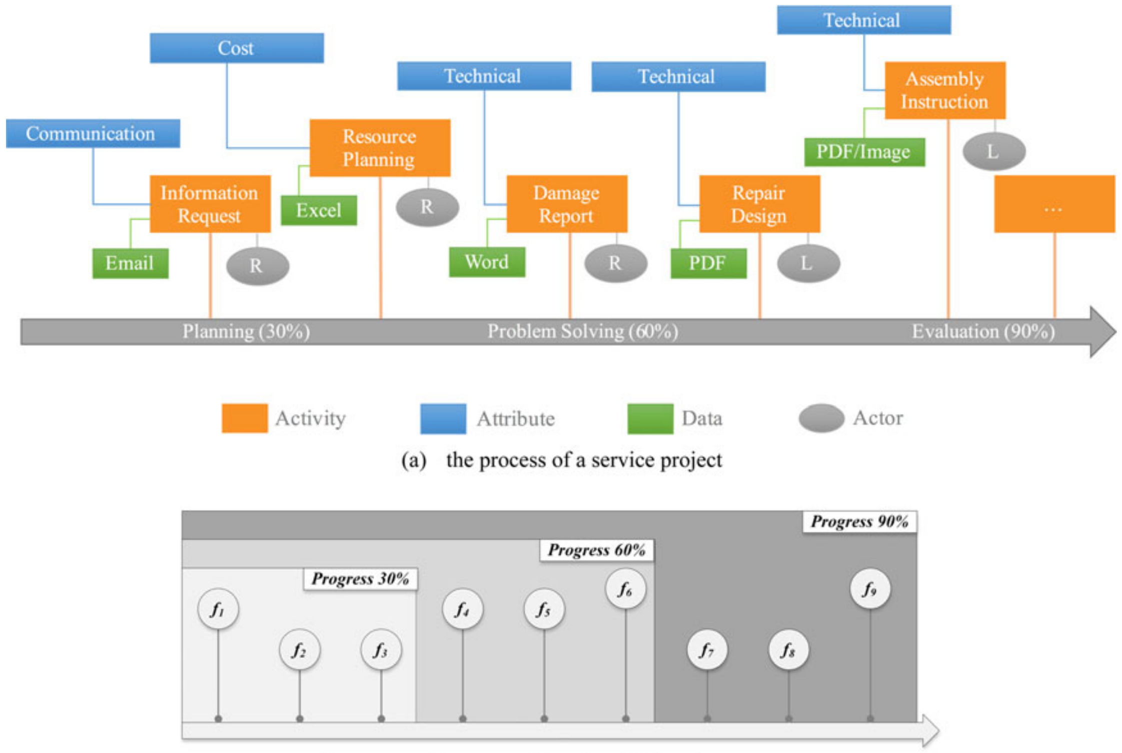
In practice, the time interval could be defined as relative value, that is, percentage, or absolute value, for example, a single day or week. It can also be defined as arbitrary value; for example, if the project stage I covers the first 10 days, the project stage II covers the following 15 days, and the project stage III covers the following 30 days, then the intervals can be defined as $\{t_1 : 10, t_2 : 25, t_3 : 55, \ldots\}$.

In the prediction process, defining the time interval considers the details of the input data. Let (τ_k, ℓ_0) denote an unlabeled feature tuple with a null characteristic label. If the timestamps of its contained features are in a time range (t_1, t_n), then this time range will be applied to define the time interval Δt, for the purpose of segmenting the training set. In the next step, the prediction process will assign a label to τ_k, according to the time-segmented training set. The detailed process is summarized as the following steps:

Step 1. For a given feature tuple τ_k, identifying its covered time range (t_1, t_n) based on its contained features.

Step 2. Determining a time interval Δt, where $\Delta t \leq (t_1, t_n)$.

Step 3. For a given training set, converting the characteristic representations into time-segmented representations by using Δt.



(a) the process of a service project

(b) segmented processes based on 30% project progress

Fig. 4. Process segmentation-based defined time interval.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

Step 4. Initializing the prediction $\ell_0 \rightarrow \tau_k$ based on the time range $t_1 + \Delta t$.

Step 5. Finding the time-segmented characteristic representations with the time range $t_1 + \Delta t$ from the time-segmented training set, that is, $p_{t_1+\Delta t} = \{c_1, \ldots, c_N\}_{t_1+\Delta t}$.

Step 6. Measuring the similarity between τ_k and each of the element contained by $p_{t_1+\Delta t}$.

Step 7. Determining ℓ_0 based on the similarity measure results, that is, $\ell_j \rightarrow \ell_0$, if $\arg\max(\text{sim}(\tau_k, c_1), \ldots, \text{sim}(\tau_k, c_N)) = \text{sim}(\tau_k, c_j)$, where $j \in [1, N]$.

According to the last step, ℓ_j is assigned as the label of τ_k; thus, the feature tuple is predicted to be correlated to the project characteristic c_j.

4. INDUSTRIAL DATA SET: AEROSPACE CASE STUDY

To evaluate the proposed model and approaches, an industrial data set captured from an aerospace manufacturer is utilized. In this case study, 156 aircraft engineering service projects are considered. Each one of them has a data repository that contains the communications, technical documents, and workflow data. The data items contain the detailed information regarding the project objectives, problem definitions, operation processes, technical solutions, and evaluations. In this case study, three tasks are included, in order to investigate the learning of pattern-related characteristics, process-related characteristics, and the learning and prediction of knowledge-related characteristics.

4.1. Learning pattern-related characteristics

In this task, the project activity level is recognized as a pattern-related characteristic. It indicates the frequency of key activities being performed during the project execution. The calculation considers the volume of activity in the workflow. In general, the activity level could indicate the level of project input or output.

In general, the feature selection process follows the five Ws (*Who*, *What*, *Where*, *Why*, and *When*) and one H (*How*) principle. For most service cases, these features are the key elements to form the characteristics. Therefore, the feature selection process is a generalized approach that suits other types of service cases with different contexts. In this scenario, the involvement of core project members and the name of key activities are considered. Such features are extracted directly from the data content using NER.

During the learning process, the feature modeling approach analyzes the project data, and then generates a feature set regarding each project stage. The feature set includes the project member names, project activities, and timestamps. Let s_k denote a project stage that covers a range of time; where $s_k = \{t_1, \ldots, t_k\}$, the feature set regarding s_k is

$$F_{s_k} = \{f_{1,t_1}, f_{2,t_2}, \ldots, f_{k,t_k}\},$$

and the vector-based representation should be

$$c_{s_k} = [w(f_{1,t_1}), \, w(f_{2,t_2}), \, \ldots, \, w(f_{k,t_k})]'.$$

The activity level regarding s_k is calculated using

$$\text{lev}(s_k) = \frac{\sum_{j=1}^{k} f_{j,t_j}}{\sum_{j=1}^{N} f_{j,t_j}},$$

where N is the total number of the features being identified. For example, if the task planning activities have been identified 6 times in the initial project stage, and the other activities have been identified 20 times at the same project stage, then the activity level regarding task planning at the initial project stage should equal to 6/(20 + 6) = 0.23.

Figure 5 shows the activity levels of two different service projects. In general, the life cycle of both projects contains four stages: the information request stage (the initial stage), requirement finalisation stage (the early to middle stage), service design stage (the middle to late stage), and evaluation stage (the late stage).

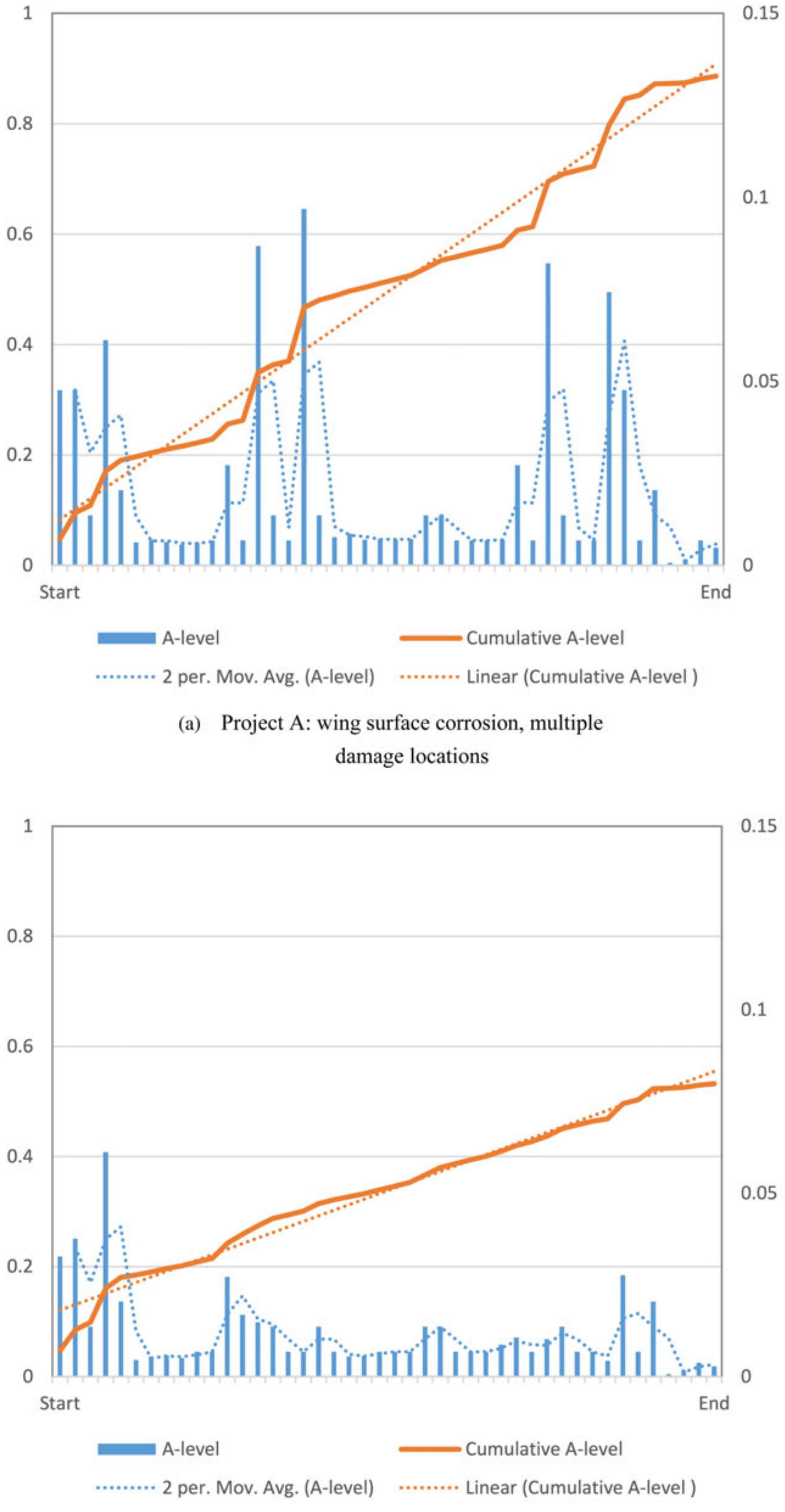
Because the projects have the same type of service requirements, that is, repairing the wing surface corrosion, similar amounts of activities regarding task planning are expected at the information request stage. According to the figure, both projects have similar activity levels at the initial stage.

Next, the cumulative activity level of Project A has a more rapid increasing trend than Project B. The reason is that Project A needs to deal with multiple corrosion locations, but Project B only deals with one. At the requirement finalization stage, Project A requires a higher level of activity than Project B, as it needs to be received and send more information. Meanwhile, at the service design stage, Project A requires a higher level of activity than Project B, as it needs to issue a higher volume of repair design solutions. At the evaluation stage, Project A again requires a higher level of activity than Project B, as it needs to perform more evaluation tasks. According to the figure, Project A has the higher activity levels at these stages than Project B.

For the reason that Project A requires more complicated engineering process and involves a higher volume of collaborations, the activity level regarding Project A could be affected more easily. According to the visualization, the change of activity level regarding Project A is clearly more frequent than Project B. This indicates the certainty regarding the execution of Project A is less than Project B, which corresponds to the facts.

Meanwhile, Project A has heavier workload than Project B; thus, it has a higher overall activity level. According to the visualization, the growth rate of activity level of Project A is clearly higher than Project B. This indicates Project A needs to complete more work than Project B within the same period of time, which also corresponds to the facts.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press



(a) Project A: wing surface corrosion, multiple damage locations

(b) Project B: wing surface corrosion, single damage location

Fig. 5. Activity levels of two service projects.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

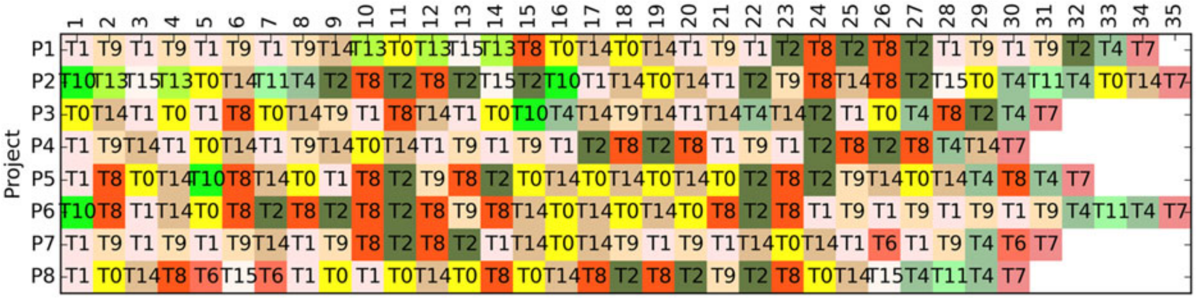


Fig. 6. Activity sequences of multiple service projects.

4.2. Learning process-related characteristics

In this task, the activity sequence is defined as a process-related characteristic. Given a project, the activity sequence is the representation of its workflow or design process. The understanding of the activity sequences regarding a type of projects could help the project members to implement the workflow standardization, design process optimization, and process knowledge reuse.

In the feature modeling process, NLP, content analysis, and modeled knowledge are used to analyze the data items on both the content and metadata levels. The activities are identified from the data content, and their activities are identified from the metadata. These activities are then organized as chronological order according to the timestamps.

Figure 6 shows the activity sequences being generated from the data set. Each row indicates the activity sequence of a project, and each *Tx* indicates an activity-related feature. For each activity sequence, its features are organized in a chronological order.

In order to compare the characteristics of different projects quantitatively, the similarity between a pair of sequences is measured by using Eqs. (6) and (7). The applied strategy is as follows: for two projects having similar characteristics, the sequence similarity between them would be high; for two projects having dissimilar characteristics, the sequence similarity between them would be low.

In this scenario, the fact is that Project P5 and P6 have the same type of service requirement, that is, corrosion damages, but Project P3 has a different one, that is, lightening damages. Therefore, the processes regarding P5 and P6 are supposed to be similar, and the processes regarding P6 and P3 should be different.

According to the learning approach, the edit distance between two sequences is equal to the number of operations to convert one to the other (see the detailed explanation in Section 3.3). For example, the number of operations (i.e., edit distance) to convert P5 to P6 (or P6 to P5) is equal to 11 (based on Eq. [6]). The edit distance is converted to a normalized similarity value with the range between 0 to 1. Based on Eq. (7), the similarity value regarding P5 and P6 is 0.656. The similarity value regarding P6 and P3 is 0.333. The results imply that P5 and P6 have similar characteristics, but P6 and P3 have different ones. In other words, the sequences of P5 and P6 have higher volume of similar patterns, but sequences of P6 and P3 have lower volume of similar patterns, which correspond to the visualizations shown in Figure 6.

4.3. Learning and predicting knowledge-related characteristics

In this task, operational complexity is defined as a knowledge-related characteristic. It covers time spending, resource consumption, and technical difficulty. Operational complexity of a service project is typically dynamic at different stages. The measure of it needs to consider the proposition of certain activities at the specific project stage. For example, the activity regarding "*issuing repair instruction*" has higher operational complexity than "*issuing technical deposition*" and "*issuing answer.*" If a project is at stage X and has higher proportion of "*repair instruction*" than another project, its operational complexity should be at a higher

Table 2. *Feature based representations for service projects*

	Incoming Message	Outgoing Message	Operator Damage Report	Stress Test	Fatigue Test	Repair & Design Approval	Technical Deposition	Repair Instruction	Answer	. . .	Total
P1	0.083	0.104	0.021	0.013	0.011	0.033	0.003	0.051	0.009	. . .	1
P2	0.053	0.063	0.017	0.008	0.005	0.029	0.043	0.000	0.013	. . .	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Status	In.	Out.	In.	Int.	Int.	Int.	Out.	Out.	Out.	. . .	. . .

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

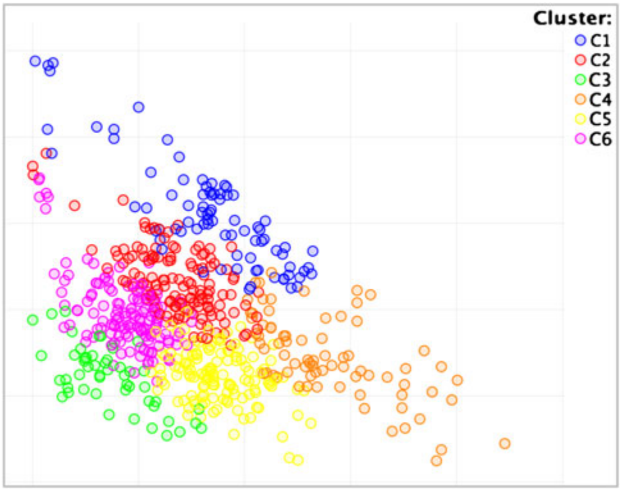


Fig. 7. The clustering result regarding operational complexity.

level than the other one. In practice, the understanding of operational complexity in an early stage could help the project members to improve the rationality of the decision making and the effectiveness of managing time and resource.

In the feature modeling process, NLP, NER, content analysis, frequency analysis, and knowledge bases are employed. The features identified from the project data include activity types, timestamps, part information, and technical terminologies. Each feature has an attribute to indicate the status of data transaction, which is identified based on the modeled knowledge. For example, *initializing fatigue test* is classified as an activity with *internal data transaction; preparing repair instruction* is classified as an activity with *outgoing transaction*; *receiving operator damage report* is classified as an activity with *incoming transaction*. The feature vector with such attributes is used as the data representation for the project (see Table 2).

In this scenario, the model employs the semisupervised learning. Clustering approach is used to create a training set from the historical project data. It categorizes the projects into multiple clusters. During the clustering process, Silhouette score is used to determine the number of clusters in an automated manner (Dudoit & Fridlyand, 2002). For the given data set, the clustering result is shown in Figure 7. The optimized cluster number is equal to six, and the projects within the same cluster have the identical color.

Table 3. *Operational complexity scores of the clusters (C1–C6)*

	Outgoing	Internal	Incoming	Complexity
C1	0.3331	0.0381	0.6288	High
C2	0.2735	0.2578	0.4687	Medium
C3	0.1497	0.6062	0.2441	High
C4	0.5551	0.1569	0.2880	Low
C5	0.3530	0.3990	0.2480	Medium
C6	0.1760	0.4290	0.3950	Medium

To understand the meaning of each cluster, the following rules contained in the knowledge bases are considered:

- The projects with high outgoing, low incoming data transactions have the low level of complexity.
- The projects with low outgoing, high incoming data transactions have the high level of complexity.
- The projects with high internal data transactions have the high level of complexity.
- The projects with balanced outgoing and incoming data transactions, and low internal data transactions, have the medium level of complexity.

Based on these rules, the level of complexity regarding each cluster is determined (see Table 3).

The projects with labels are treated as the training data, which will be used to predict the operational complexity of unlabeled projects. To evaluate the performance of the characteristic prediction process, 10-fold cross-validation is used. The set of labeled projects is divided into a training set (90% of the data set) and a test set (10% of the data set). During the evaluation process, the prediction algorithms, including support vector machine, artificial neural network, and random forest, are tested with the model respectively. In this evaluation, F score is used as the prediction performance measure.

In the test set, each project has a label ℓ_L being assigned by the learning process previously. Such a label is considered as the ground truth, and it is invisible to the predictors. In the training set, the label of each project is visible to the predictors. The predictors need to determine what label should be assigned to the unlabeled projects in the test set based on labeled projects in the training set. The project representations in both test set and training set are segmented by using normalized time intervals, which are 0%–30%, 0%–50%, 0%–70%, and 0%–90%. According to the training set, the prediction process will assign a label to each test data, for all the stages, for example, $\ell_1 \rightarrow$ 0%–30%, $\ell_2 \rightarrow$ 0%–50%, $\ell_3 \rightarrow$ 0%–70%, and $\ell_4 \rightarrow$ 90%. The evaluation process is then to compare whether the assigned labels $\{\ell_1, \ell_2, \ell_3, \ell_4\}$ are identical to the ground truth ℓ_L. The F score of this prediction processes is shown in Table 4.

As shown in the table, the artificial neural network based model has the best performance with 0%–30% and 0%–

Table 4. *The F score of the prediction model*

	0%–30%	0%–50%	0%–70%	0%–90%
SVM based	0.712	0.745	0.802	0.909
ANN based	0.732	0.761	0.844	0.928
RF based	0.725	0.773	0.848	0.917
Average	0.723	0.760	0.831	0.918

Note: SVM, Support vector machine; ANN, artificial neural network; RF, random forest.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

90% time intervals, and the random forest based model has the best performance with 0%–50% and 0%–70% time intervals. In general, the prediction performance is proportional to the completion level of projects. It implies that the more data is considered by the model, the better prediction performance it will be. By using the model, the operational complexity of a service project could be predicted with a 70+% accuracy by using 30% of the project data.

5. CONCLUSIONS AND FUTURE WORK

The engineering service project needs to take into account various types of information such as the product features, functionalities, and reliabilities, as well as the customer requirements, market demands, and organizational context. Consequently, a large amount of data and heterogeneous digital objects are generated, utilized, and shared during the design process. The information overload prevents the effective reuse of project data and knowledge, making the understanding of project characteristics difficult.

In order to improve the characteristic understanding process, the requirements such as reducing the human intervention, utilizing the collective knowledge, and developing the data-driven models are critical. It implies that the integration of data mining and machine learning techniques with the understanding process is necessary.

The work presented in this paper proposed a characteristic learning and prediction model and data analytical approaches based on the techniques including natural language processing, named entity recognition, content analysis, sequence analysis, feature modeling, data clustering, and classification. The learning process of the model identifies the pattern-related, sequence-related, and knowledge-related characteristics from the project data. Furthermore, the prediction process of the model predicts the characteristics of projects at their early stages, based on the incomplete data.

The case study using an industrial data set shows the proposed model and approaches have the capability to learn the characteristics of collaborative engineering service projects, and also to predict certain characteristics of the projects based on incomplete data. Future work includes the test of the model with different data sets, the improvement of learning, and predication processes by integrating various knowledge resources.

ACKNOWLEDGMENTS

The research reported in this paper is funded by Engineering and Physical Sciences Research Council Grant EP/K014196/1. The authors thank the industrial collaborators and their engineers for their input and support on this project.

REFERENCES

Ahmad, S., Mallick, D.N., & Schroeder, R.G. (2013). New product development: impact of project characteristics and development practices on performance. *Journal of Product Innovation Management 30(2)*, 331–348.

Baines, T.S., Lightfoot, H.W., Evans, S., Neely, A., Greenough, R., Peppard, J., et al. (2007). State-of-the-art in product-service systems. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 221(10)*, 1543–1552.

Chandrasegaran, S.K., Ramani, K., Sriram, R.D., Horváth, I., Bernard, A., Harik, R.F., & Gao, W. (2013). The evolution, challenges, and future of knowledge representation in product design systems. *Computer-Aided Design 45(2)*, 204–228.

Chen, H., Chiang, R.H., & Storey, V.C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly 36(4)*, 1165–1188.

Cho, K., Hong, T., & Hyun, C. (2009). Effect of project characteristics on project performance in construction projects based on structural equation model. *Expert Systems With Applications 36(7)*, 10461–10470.

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.

Chuang, P.-T. (2007). Combining service blueprint and FMEA for service design. *Service Industries Journal 27(2)*, 91–104.

Chungoora, N., Young, R.I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N.A., et al. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. *Computers in Industry 64(4)*, 392–401.

Doultsinou, A., Roy, R., Baxter, D., Gao, J., & Mann, A. (2009). Developing a service knowledge reuse framework for engineering design. *Journal of Engineering Design 20(4)*, 389–411.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a data set. *Genome Biology 3(7)*, 1–21.

Engwall, M., & Jerbrant, A. (2003). The resource allocation syndrome: the prime challenge of multi-project management? *International Journal of Project Management 21(6)*, 403–409.

Feng, G., Cui, D., Wang, C., & Yu, J. (2009). Integrated data management in complex product collaborative design. *Computers in Industry 60(1)*, 48–63.

Gheyas, I.A., & Smith, L.S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition 43(1)*, 5–13.

Goh, Y.M., Newnes, L., Settanni, E., Thenent, N., & Parry, G. (2015). Addressing uncertainty in estimating the cost for a product-service-system delivering availability: epistemology and ontology. In *Ontology Modeling in Physical Asset Integrity Management* (Ebrahimipour, V., & Yacout, S., Eds.), pp. 199–219. Cham, Switzerland: Springer.

Gopsill, J., Jones, S., Snider, C., Shi, L., McMahon, C., & Hicks, B. (2014). Understanding the engineering design process through the evolution of engineering digital objects. *Proc. 13th Int. Design Conf.* (*DESIGN 2014*), Dubrovnik, Croatia, May 19–May 22.

Griffin, A. (1997). The effect of project and process characteristics on product development cycle time. *Journal of Marketing Research 34(1)*, 24–35.

Harding, J., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering 128(4)*, 969–976.

Jones, S., Payne, S., Hicks, B., Gopsill, J., Snider, C., & Shi, L. (2015). Subject lines as sensors: co-word analysis of email to support the management of collaborative engineering work. *Proc. Int. Conf. Engineering Design 2015 (ICED 2015).* Milan, Italy, July 27–30.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems With Applications 40(4)*, 1034–1045.

Kankar, P.K., Sharma, S.C., & Harsha, S.P. (2011). Fault diagnosis of ball bearings using machine learning methods. *Expert Systems With Applications 38(3)*, 1876–1886.

Köksal, G., Batmaz, İ., & Testik, M.C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems With Applications 38(10)*, 13448–13467.

Li, C., McMahon, C., & Newnes, L. (2009). Annotation in product lifecycle management: a review of approaches. *Proc. ASME 2009 Int. Design Engineering Technical Conf./Computers and Information in Engineering Conf.*, pp. 797–806. New York: American Society of Mechanical Engineers.

Li, Y.-F., Xie, M., & Goh, T.N. (2009). A study of project selection and feature weighting for analogy based software cost estimation. *Journal of Systems and Software 82(2)*, 241–252.

Luchs, M., & Swan, K.S. (2011). Perspective: the emergence of product design as a field of marketing inquiry. *Journal of Product Innovation Management 28(3)*, 327–345.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press

Meredith, J.R., & Mantel, S.J., Jr. (2011). *Project Management: A Managerial Approach*. Hoboken, NJ: Wiley.

Mesihovic, S., Malmqvist, J., & Pikosz, P. (2004). Product data management system-based support for engineering project management. *Journal of Engineering Design 15(4)*, 389–403.

Paroutis, S., & Al Saleh, A. (2009). Determinants of knowledge sharing using Web 2.0 technologies. *Journal of Knowledge Management 13(4)*, 52–63.

Pascal, A., Thomas, C., & Romme, A.G.L. (2013). Developing a human-centred and science-based approach to design: the knowledge management platform project. *British Journal of Management 24(2)*, 264–280.

Petersen, K.J., Handfield, R.B., & Ragatz, G.L. (2003). A model of supplier integration into new product development. *Journal of Product Innovation Management 20(4)*, 284–299.

Rudin, C., Waltz, D., Anderson, R.N., Boulanger, A., Salleb-Aouissi, A., Chow, M., et al. (2012). Machine learning for the New York City power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34(2)*, 328–345.

Sahin, F., Yavuz, M.Ç., Arnavut, Z., & Uluyol, Ö. (2007). Fault diagnosis for airplane engines using Bayesian networks and distributed particle swarm optimization. *Parallel Computing 33(2)*, 124–143.

Settanni, E., Thenent, N.E., Newnes, L.B., Parry, G., & Goh, Y.M. (2015). To cost an elephant: an exploratory survey on cost estimating practice in the light of product-service-systems. *Journal of Cost Analysis and Parametrics 8(1)*, 1–22.

Shi, L., Gopsill, J., Newnes, L., & Culley, S. (2014). A sequence-based approach to analysing and representing engineering project normality. *Proc. IEEE 26th Int. Conf. Tools With Artificial Intelligence (ICTAI)*, pp. 967–973. Washington, DC: IEEE Computer Society.

Shi, L., Gopsill, J., Snider, C., Jones, S., Newnes, L., & Culley, S. (2014). Towards identifying pattern in engineering documents to aid project planning. *Proc. 13th Int. Design Conf.* (*DESIGN 2014*), Dubrovnik, Croatia, May 19–May 22.

Shi, L., Newnes, L., Culley, S., & Snide, C. (2015). Process reconstruction and visualisation for collaborative engineering projects. *Proc. 13th Int. Conf. Manufacturing Research*, Bath, UK, September 8–10.

Shi, L., & Setchi, R. (2013). Enhanced semantic representation for improved ontology-based information retrieval. *International Journal of Knowledge-Based and Intelligent Engineering Systems 17(2)*, 127–136.

Snider, C., Jones, S., Gopsill, J., Shi, L., & Hicks, B. (2014). A framework for the development of characteristic signatures of engineering projects. *Proc. 13th Int. Design Conf.* (*DESIGN 2014*), Dubrovnik, Croatia, May 19–May 22.

Wagstaff, K.L. (2012). Machine learning that matters. *Proc. 29th Int. Conf. Machine Learning (ICML-12)*, Edinburgh, Scotland, June 26–July 1.

Walter, G. (2014). Determining the local acceptance of wind energy projects in Switzerland: the importance of general attitudes and project characteristics. *Energy Research & Social Science 4*, 78–88.

Wang, K. (2007). Applying data mining to manufacturing: the nature and implications. *Journal of Intelligent Manufacturing 18(4)*, 487–495.

Wasiak, J., Hicks, B., Newnes, L., Loftus, C., Dong, A., & Burrow, L. (2011). Managing by e-mail: what e-mail can do for engineering project management. *IEEE Transactions on Engineering Management 58(3)*, 445–456.

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing 21(6)*, 2560–2574.

Wu, D., Thames, J.L., Rosen, D.W., & Schaefer, D. (2012). Towards a cloud-based design and manufacturing paradigm: looking backward, looking forward. *Proc. ASME 2012 Int. Design Engineering Technical Conf./Computers and Information in Engineering Conf.*, pp. 315–328. New York: American Society of Mechanical Engineers.

Zhang, D., Hu, D., Xu, Y., & Zhang, H. (2012). A framework for design knowledge management and reuse for product-service systems in construction machinery industry. *Computers in Industry 63(4)*, 328–337.

Zhen, L., Jiang, Z., & Song, H.-T. (2011). Distributed knowledge sharing for collaborative product development. *International Journal of Production Research 49(10)*, 2959–2976.

Lei Shi is a Research Officer in the Department of Mechanical Engineering at the University of Bath. His research focuses on cloud-based design and manufacturing, enterprise data mining, machine learning, complex/knowledge-based systems design, computational modeling, and human–computer interaction. He has published more than 25 scientific papers in peer-reviewed journals and conferences.

Linda Newnes is a Professor in the Department of Mechanical Engineering at the University of Bath. Her research focuses on through life costing from concept design through the in-service/in-use phases. The sectors/application areas for her activities include, for example, aerospace, defense, medical device design, oil and gas, green technologies, modeling uncertainty in through life costing, modeling in-service costs, and trade-off analysis between specification and cost.

Steve Culley is a Professor in the Department of Mechanical Engineering at the University of Bath. He is Head of Design and Manufacturing, having worked in the steel, fluid power, and rubber industries. Steve's research is in the engineering design field, focusing on the provision of information and knowledge to support engineering designers.

Bruce Allen is a Principal Wing Design Repair Engineer at Airbus. He studied engineering at Manchester University, graduating with a BS Honours Degree, and launched his engineering career with BAE Systems in 1987. As a conceptual engineer he worked on several successful programs, culminating in a joint project for the US and Saudi governments. At Airbus he is responsible for all technical aspects of aircraft wing repair designs produced for in-service aircraft by a multinational team located across the world. Airbus presented new experiences, working with many different aircraft, developing new proficiencies and skills over several years, which led naturally to the current role. This role embraces his many years of engineering experience delivering benchmark, innovative solutions to complex problems.

https://doi.org/10.1017/S0890060416000470 Published online by Cambridge University Press