Iterative evaluation and threshold rules establishment.

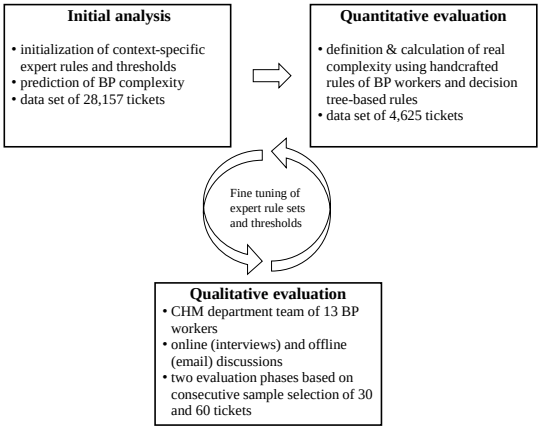

Fig. 1. Iterative evaluation and threshold rules establishment process

The **initial analysis** (see the box "*Initial analysis*" in Fig. 1) of the iterative evaluation strategy was carried out to set up initial expert rules and thresholds. First, data processing took place. The first data set was processed and converted into a CSV-formatted text corpus with more than 1,000,000 documents (text entries) of English, German, and English-German ticket texts created from 2015 to 2018. After removing duplicates and selecting prevailingly English texts (tickets with more than 80% of English words), the data sample comprised 28,157 entries. Pre-processing and extraction of the knowledge types were conducted using Python 3.4. The development of handcrafted rules and thresholds and quantitative evaluation were implemented iteratively using Microsoft Office Excel 2016.

In this ***first step***, the objective, subjective, and meta-knowledge were subsequently extracted using the data set of 28,157 tickets.

In the ***second step***, to evaluate the initial results of the first step, 30 randomly selected tickets with predicted RfC complexity values were presented to the CHM department workers. This first evaluation round included an introductory evaluation questionnaire, interview, and offline follow-up sessions conducted in May 2019. The interview was divided into three parts. First, we introduced the objectives of the interview, research motivations, theoretical and methodological background. Second, the method of knowledge extraction and BP complexity prediction was illustratively presented using a sample of 30 tickets. The discussion of the discrepancies between predicted RfC complexity and historical data-based complexity was shifted to offline (email) due to additional information from the IT ticketing system necessary for estimation. In the end of the interview, a Q&A session was conducted following a so-called funnel model [1]. We started with open questions and moved towards more specific ones regarding possible practical implications of the complexity prediction. The results evidenced the following positive feedback: (i) recommendations in the form of templates and relevant historical RfC ticket data, (ii) the prioritization of an incoming RfC ticket in the form of a dashboard for the correct time and workforce management in the team, (iii) automatic filling in of the ticket complexity field in the IT ticketing system.

The offline discussion of the RfC complexity values yielded the results presented in Table 1, point 1. The table gives an overview of qualitative and quantitative evaluation steps and statistics. Overall precision is the relative number of correctly predicted RfC complexity values compared to the historical data based RfC complexity. Recalls are calculated for predicted RfC complexity and represent a fraction of relevant values that have been retrieved over the total amount of relevant values.

Table 1. Complete iterative evaluation process and statistics of RfC complexity prediction

	low	medium	high
1. Qualitative evaluation with subject matter experts based on 30 tickets from the 1st data set of 28,157 tickets			
Recall	69%	55%	67%
Overall precision	63%		
2. Quantitative historical data-based evaluation of 2nd data set of 4,625 tickets (1st round)			
Recall	51.2%	28.4%	9.1%
Overall precision	45%		
3. *Qualitative evaluation with subject matter experts based on 60 tickets from the 2nd data set of 4,625 tickets*			
Recall	62%	10%	0%
Overall precision	54%		
4. *a. Quantitative historical data-based evaluation of 1st data set of 28,157 tickets (2nd round)*			
Recall	70%	69%	50%
Overall precision	65%		
4. *b. Quantitative historical data-based evaluation of 2nd data set of 4,625 tickets (2nd round)*			
Recall	73.9%	71.9%	40.7%
Overall precision	61.75%		
5. *Quantitative historical data-based evaluation of 2nd data asset of 4,625 tickets using CART based rules*			

Recall	75.6%	61.6%	50.2%
Overall precision	62.27%		

As a result of this second step, we collected the following findings for **qualitative improvements**: (i) enrichment of the DML Taxonomy and BS Lexicon with the one- and bi-grams indicating simple and complex problem solving and (ii) development of the handcrafted rules and thresholds, identification of the necessary historical ticket data to calculate the (real) RfC complexity to be used in quantitative evaluation. Hence, we amended the DML Taxonomy and BS Lexicon vocabularies with such one- and bigrams as "(no, not) affected", "(no) PSO" (Projected Service Outage), "(no) impact", "(no, short, zero) downtime", "test", "(no, not) production", "(no, not) prod". We also added German equivalents of such adverbs as "no", "not", i.e. "kein(e)", "nicht", for the case of English-German RfC ticket texts. To identify the historical data based RfC complexity, the following rules based on the historical data were set: (i) the presence of the mentioned one- and bi-grams in the IT ticketing system fields "Impact description" and "Brief description" of the ticket (RegEx based free text search); (ii) the number of tasks per ticket (count of tasks, integer data type); (iii) the number of configuration items, specifically applications involved in the RfC ticket (count of applications, integer data type); 4) risk type of ticket (enumeration, ordinal scale of "low", "medium", "high").

As the ***third step***, we obtained a new data set including the historical data necessary to identify the historical data based RfC complexity. The data set represented a CSV-formatted file with around 8,643 entries in English, German, and English-German. Following the same process as with the first data set, we selected prevailingly English texts (tickets with more than 80% of English words). Hence, the final data set included 4,625 entries. The period of the RfCs was from January to May 2019. The three knowledge types, i.e., DML, BS, and Readability, were extracted from the pre-processed unstructured RfC texts, followed by the prediction of RfC complexity and identification of the historical data based RfC complexity using the provided historical data and threshold rules discussed with the subject matter experts, i.e., CHM department workers. However, the quantitative evaluation results were not satisfying. The overall precision totals approximately 45%. The predicted RfC complexity recalls make up 51.2% low, 28.4% medium, and only 9.1% high correspondingly (see Table 1, point 2).

Therefore, in the ***fourth step***, in July 2019, we conducted another qualitative interview round in an offline (email) mode. For this purpose, 60 randomly selected tickets with both predicted (textual data based) and historical data based RfC complexity values were presented to the CHM department workers (see the box "**Quantitative evaluation**" in Fig. 1). The sample contained 54% correctly and 46% incorrectly classified complexities of low, medium, and high values. The goal was to adjust the threshold rules for the historical data based RfC complexity. In the offline discussions, the cases of discrepancies between identified historical data based RfC complexity and the one assigned by the experts were reviewed (for the evaluation statistics, see Table 1, point 3). Using (i) the modified rules regarding keywords and (ii) threshold rules for the historical data based RfC complexity, we adjusted also our thresholds for DML, BS, Readability, and BP complexity. Afterward, we performed another quantitative evaluation round on both data sets. As the results show, we achieved an improvement in (i) prediction and (ii) evaluation showing the overall precision of 65% and approximately 62% for two data set correspondingly (see Table 1, point 4).

In addition to the handcrafted rules the historical data based RfC complexity, we applied a technology-based approach – the recursive partitioning classification and regression trees (CART) method [2] with complexity parameter cp=0.056 and measures of the error in classification xerror=0.39. We used the mentioned set of manually evaluated 60 tickets as a training sample and a data set of 4,625 tickets as a test sample. The results can be seen in Table 1, point 5.

[1] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empir. Softw. Eng. 14 (2009) 131–164. https://doi.org/10.1007/s10664-008-9102-8.

[2] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, Decision Trees: An Overview and Their Use in Medicine - Semantic Scholar, J. Med. Syst. . 26 (2002) 445–463. https://doi.org/10.1023/A:1016409317640.