EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Intelligent radar software defect classification approach based on the latent Dirichlet allocation topic model

Xi Liu[1], Yongfeng Yin[2*], Haifeng Li[3], Jiabin Chen[1], Chang Liu[3], Shengli Wang[1] and Rui Yin[2]

* Correspondence: yyf@buaa.edu.cn
[2]School of Software, Beihang University, Beijing 100191, China
Full list of author information is available at the end of the article

Abstract

Existing software intelligent defect classification approaches do not consider radar characters and prior statistics information. Thus, when applying these appaoraches into radar software testing and validation, the precision rate and recall rate of defect classification are poor and have effect on the reuse effectiveness of software defects. To solve this problem, a new intelligent defect classification approach based on the latent Dirichlet allocation (LDA) topic model is proposed for radar software in this paper. The proposed approach includes the defect text segmentation algorithm based on the dictionary of radar domain, the modified LDA model combining radar software requirement, and the top acquisition and classification approach of radar software defect based on the modified LDA model. The proposed approach is applied on the typical radar software defects to validate the effectiveness and applicability. The application results illustrate that the prediction precison rate and recall rate of the poposed approach are improved up to 15 ~ 20% compared with the other defect classification approaches. Thus, the proposed approach can be applied in the segmentation and classification of radar software defects effectively to improve the identifying adequacy of the defects in radar software.

Keywords: Radar software, Software defect, Defect classification, Latent Dirichlet allocation (LDA) topic model

Radar equipment is responsible for homeland air defense, detection and perception, space attack and defense, formation coordination, and other important tasks. Thus, it is very important for the monitoring and management, anti-missile self-defense in the field of the land, sea, and air, as well as defeating the enemy on the battlefield. With the rapid development of digitalization, networking, and intelligence of weapons and equipment, software has become the core component of radar system. The core functions such as target detection, data analysis, real-time processing, and equipment monitoring in radar, as well as the important tasks such as search, tracking, identification, anti-jamming, and so on, and the performance improvement of rapid adaptation to combat environment are all realized by software. Once the failure of software occurs, it may result in the mission fails, or the equipment damage or even casualties. Thus,

software quality has become a key element which has important effection on the quality of radar system.

At the same time, the radar software can be characterized by high real-time, complex task scenarios, high functional integration, frequent data interaction, and often run on specially designed boards or chips. Usually, airborne radars contain more than 100,000 lines of C/C++ and FPGA code, early warning radars even include more than a million lines of C/C++ and FPGA code, large radar systems often contain more than five display and control terminals, and there are hundreds of internal and external control instructions, real-time processing data volume of more than 1Gbps. These reasons lead to the complex mechanism of radar software defects; how to accurately identify and predict the distribution of potential defects and weak links has become the key to affect the efficiency of radar software testing and equipment quality.

Classification techniques for software defects are an effective way to improve the identification and prediction of defects [1]. Most of software defect data are described by the natural language texts with irregularities and duality. It is difficult for computers to effectively handle and classify the data of software defects. If we only rely on manual means to classify the historical defect data and reuse the same category of software defects, it requires more workload and is affected by human subjective factors, which is easy to produce omissions and difficult to guarantee the efficiency and quality of defect classification. Therefore, one of the effective ways to solve this problem is that classifying and reusing software defects intelligently by the artificial intelligence technologies such as natural language processing.

Compared with general software defects, radar software defects have more complex domain characteristics (e.g., GJB 4429 and 5090, which list a large number of terms specific to the radar domain), and it is difficult to perform accurate text disambiguation or topic acquisition on radar software defect data with currently available text disambiguation techniques (e.g., conditional random field model) or general topic acquisition models (e.g., latent Dirichlet allocation topic model, namely latent Dirichlet distribution topic model, the following referred to as LDA topic model). In addition, most of the existing data classification techniques based on topic models are unsupervised learning techniques with less consideration of the a priori statistical features of the data, and the accuracy and recall of the classification of radar defect data are affected. At the same time, radar users have high requirements for the overall functional performance of the equipment and other quality factors, while software testing generally focuses on whether the whole machine and subsystems are compatible with the requirements, as well as the reliability of software tasks and the potential needs of user experience, and requires a lot of human resources and other resources for black-box software testing at the system and subsystem levels. Therefore, to improve the efficiency of radar software testing and reduce the testing workload, it is necessary to study the defect classification methods for radar software requirements, applicable to system testing and configuration item testing.

Therefore, this project proposes a defect classification method for radar software based on an improved LDA topic model (the following referred to as RadarDCP). Firstly, we propose the radar domain dictionary to realize the accurate word classification of radar software defects. And then, we propose the modified LDA topic model incorporating the features of radar software requirements to obtain the potential topics

of software requirement such as the function names and the interface names, to improve the accuracy of the LDA model topic identification; then, based on the LDA topic model, we realize radar software defect classification. Finally, we propose the experiment application on the typical radar software defects to validate the effectiveness and usability of the apporach proposed in this paper.

The structure of this paper is organized as follows: Section 1 analyzes the existing research on software defect prediction and gives a framework of intelligent prediction techniques for software defects, Section 2 proposes an improved LDA topic model incorporating the features of radar requirements, Section 3 proposes the topic acquisition and classification of radar defect data based on the improved LDA model, Section 4 presents an experimental study of the proposed approach based on the typical software defect data of radar systems, and Section 5 gives the conclusion.

1 Existing work analysis and technical framework

1.1 Related works

At present, the research on the application of artificial intelligence techniques in the domain of software defect classification and prediction is divided into two main types as follows:

1.1.1 (1) Intelligent algorithm-based software testing usage [2–6]

At present, the existing software testing knowledge reuse technologies based on the intelligent algorithm are mostly focused on the research of test case reuse, in another word, based on the historical test cases, with the help of an intelligent algorithm, and the reusable test cases are recommended for the current projects to improve software testing efficiency and reduce work costs. This includes the reuse of software test cases based on document similarity, the classification and intelligent retrieval of reusable test cases, the automatic generation of software failure modes, etc. Most of the existing techniques are based on the attribute metric of the software product, the distribution of defects, the number of defects, and other information for prediction. However, the problem to be solved in this paper is the reuse prediction of radar software defect data, namely how to predict similar defect problems from the existing historical defect data sets based on the current radar software requirements. The existing software testing knowledge reuse techniques based on intelligent algorithms are mostly focused on test case reuse, which does not effectively consider the characteristics and failure mechanism of radar software. Thus, it is difficult to effectively achieve the intelligent prediction capability of radar software defects.

1.1.2 (2) Data text splitting technology

Text disambiguation is a prerequisite for accurate classification and prediction of software defect data. At present, the more common text separation techniques include lexicon-based matching [7], i.e., based on the constructed dictionary (e.g., the modern Chinese dictionary, etc.), the “data text” can be sliced according to the certain rules by the segmentation algorithms such as inverse maximum matching and forward maximum matching. The statistical model-based word separation method [8], which transforms the text separation problem into a sequence annotation problem, is implemented

with the help of statistical models for text separation such as the hidden Markov model, conditional random field model (CRF), and maximum entropy model. Deep learning-based word separation methods [9], namely deep learning algorithms such as convolutional neural networks (CNN), recurrent neural networks (RNN), are used to learn from the annotated training set in order to achieve text separation in terms of word frequency, contextual relationships, etc. The radar software test data usually contains a lot of technical terms, such as silent zone, sector, and identification zone. It is still difficult to accurately classify the terminology in the radar domain, for example, silent zone may be divided into two meaningless words, silent and zone.

1.1.3 (3) Topic model-based text classification techniques

The main idea of the topic model is that a text is considered to be composed by two types of structures: document-topic and topic-word, i.e., a document is a probability distribution of several topics. At the same time, each topic is a probability distribution of words. "Keyword Extraction based on Topic Models" is a technique of training and learning from historical defective data (i.e., training set) by the Dirichlet allocation and plain Bayesian model, and decomposing the defective data text into multiple topic models, each of which consists of one or more words that characterize the keywords of a topic of the text. The most common topic models are the PLSA model, LSA model, LDA model, etc. Based on the acquired topic models, the classification of software defect data can be achieved.

Among these topic models, the LDA topic model does not require additional annotation and processing of the training set, is unsupervised learning, has less technical difficulty and workload, and has been more widely studied and applied in text classification [10–12]. However, the current LDA topic model has less consideration for the demand features and a priori statistical features of the defect data, and there are problems such as the forced assignment of implicit topics, poor integration of classification results with demand, and lack of easy interpretation, which affect the accuracy and recall rate of radar defect data classification.

1.2 Background

In this paper, existing techniques such as the LDA topic model and inverse maximum matching algorithm are required. In this section, the basic principles of these existing techniques are explained as follows:

1.2.1 (1) Principle of LDA topic model

The LDA topic model treats the text as consisting of a multi-layer architecture of document-topic and topic-word. Each document can be viewed as a probability distribution of several topics; at the same time, each topic can be viewed as a probability distribution of several words. The core is the Bayesian estimation process of computing the posterior topic distribution of documents based on the Dirichlet prior hypothesis of document topic distribution and topic word distribution, combined with the corpus. After model inference and parameter estimation, the text corpus is decomposed into multiple topic vectors, and each topic vector consists of one or more words, which are

characterized by the certain topic of the text. The specific principle of the LDA topic model is shown in the following Fig. 1 [11]:

In Fig. 1, assuming that the document corpus (for example, radar software defect data) has D documents, there are N words in the corpus, $W_{d,n}$ represents the nth word in the dth document, and each document consists of k topics Composition, the topic-word probability distribution under each topic ϕ_k obeys the Dirichlet distribution with β as the parameter, θ_d is the document-topic distribution, each document corresponds to a different topic distribution, θ_d obeys the Dirichlet distribution withα as the parameter, $Z_{d,n}$,n represents the specified distribution between topics and words within the defect data d, and $Z_{\mathrm{d,n}}$ obeys the polynomial distribution with θ_z as the parameter.

1.2.2 (2) The reverse maximum matching algorithm

The maximal matching algorithm is the main algorithm applied to text separation, which includes forward maximal matching algorithm, reverse maximal matching algorithm, and two-way matching algorithm. The main principle is to cut out a single word string, and then compare it with the lexicon, if it is a word, record it, otherwise continue the comparison by adding or subtracting a single word, and terminate if there is still a single word left, or treat it as unregistered if the single word string cannot be cut.

The reverse maximum matching method is usually abbreviated as the RMM method, which starts from the end of the processed document to match and scan, and each time takes the last 2i characters (i character string) as the matching field. If the matching fails, remove the top of the matching field. Correspondingly, the text segmentation dictionary is a reverse order dictionary, in which each entry will be stored in reverse order. In actual processing, the document is first processed in reverse order to generate reverse order documents. Then, according to the reverse order dictionary, the forward maximum matching method can be used to process the reverse order document.

1.3 Framework for intelligent classification techniques for radar software defects

This paper proposes an intelligent classification method for radar software defects based on an improved LDA topic model (referred to as RadarDCP), whose technical framework is shown in Fig. 2. Based on the contents of Fig. 2, the overall technical scheme of this paper is described as follows:

Input:

- Historical software defect data set (test problem reports, the FMEA lists, etc.)
- The new project software requirements (such as the function name, interface name, etc.)

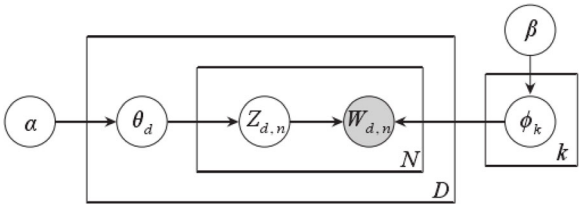

Fig. 1 The latent Dirichlet allocation topic model

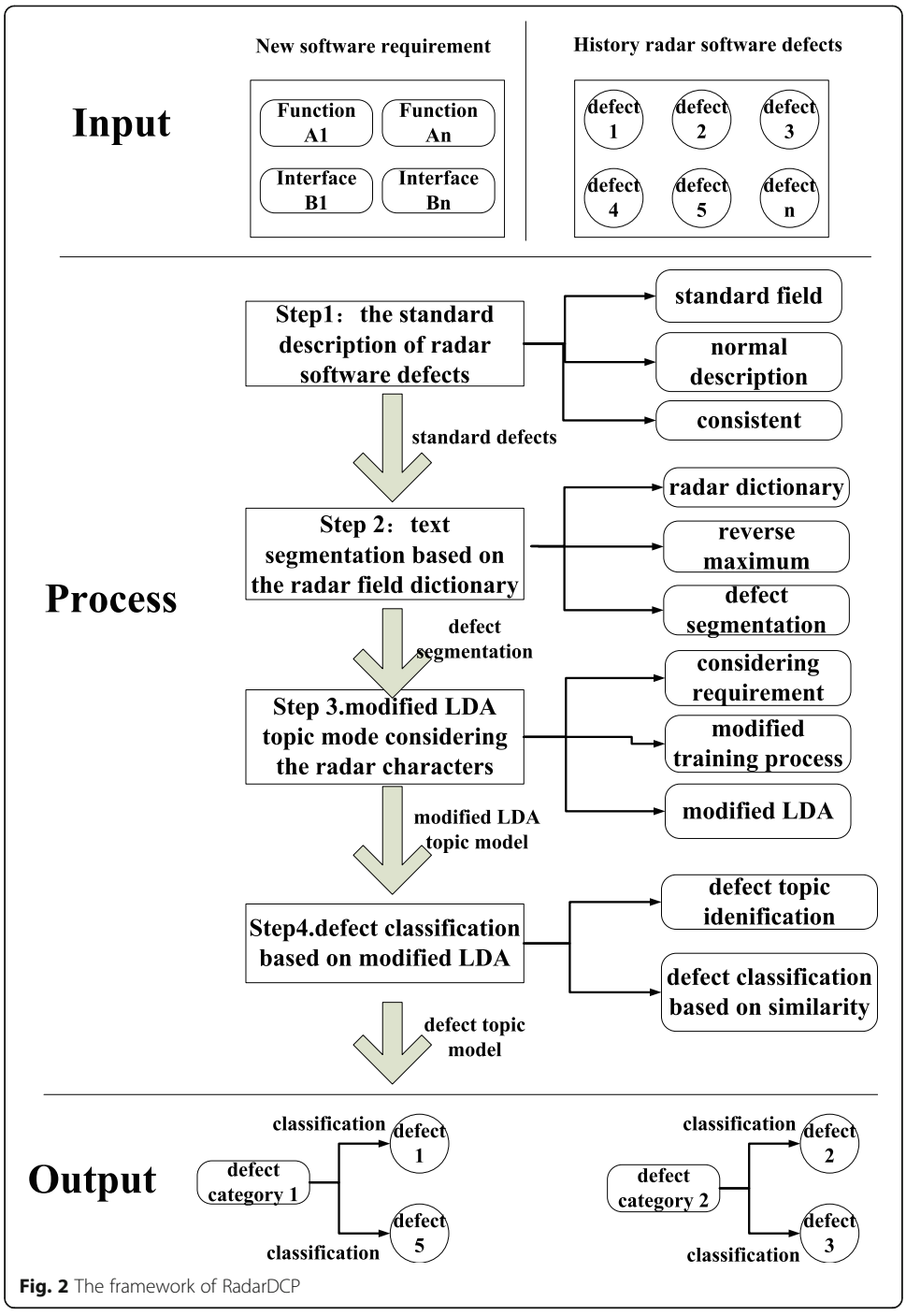

Fig. 2 The framework of RadarDCP

Process:

Step 1: Standardized preprocessing of radar software defect data

Based on the mechanism of software defect generation and propagation, a standardized software defect data record structure is developed, including related functions, defect cause, defect description, defect impact, impact level, control measures, and the other fields. At the same time, for each field content description, grammatical structure,

consistency, etc. to ensure that the data format and content without duplication, to form an iterative reusable radar software defect data set.

Step 2: Segmentation of defective text based on radar domain dictionary

Analyze the radar domain standards, demand documents, historical test data, and other corpus sets, and build the radar domain dictionary from multiple perspectives such as professional terms, synonyms, demand information, stop words, and abnormal types. Then, with the help of a reverse maximum matching segmentation algorithm, we can achieve the accurate text segmentation of radar software defect data text.

Step 3: Modified LDA topic model incorporating radar demand characteristics

The traditional LDA topic model is an unsupervised learning process, which suffers from the problem of forced assignment of implied topics, i.e., it is impossible to control the category and direction of topic acquisition for defective data, which may lead to uninterpretable classification results of defective data. In this paper, we propose an improved LDA topic model incorporating radar requirement features by referring to the Labeled-LDA method [13], in which the required elements such as the name of radar software function, interface data, interface type, etc., are used as extended features and incorporated into the LDA model learning process to adjust the parameter estimation of the distribution function. Correspondingly, the LDA model learning results are guided to obtain a topic model for radar requirements features.

Step 4: Radar defect data topic acquisition and classification based on improved LDA model

Based on the improved LDA topic model considering the radar requirement features, the radar software historical defect data are trained and learned to form multiple defect topic models. Based on the correlation between each defect data and the topic models, the defect data are classified according to the topic models, and the set of keywords for each topic model is obtained.

Output: New project software requirements (function name, interface name, etc.) predicted possible defect data

2 Modified LDA topic model incorporating radar demand characteristics

When applying the LDA model to radar software defect data, the following problems exist: (1) it is difficult to achieve accurate terminology classification of radar domain terms, such as silent zone, target identification zone, and sector, (2) unsupervised learning process, the obtained topic model is difficult to interpret, and it is difficult to achieve radar software defect data classification according to the software requirements of new projects. To address these two problems, this paper carries out an improved LDA topic model incorporating radar requirement features as follows:

2.1 Defect data text segmentation based on radar domain dictionary

Firstly, we summarize the radar domain-specific terminology collection for the standard specification, system requirement design, historical test questionnaire, use case list, and another corpus in the radar domain. Then, the radar domain-specific terms are analyzed, and the radar domain dictionary is built by determining the set of synonyms for each term with the experience of experts in radar domain system design or software development. This paper considers the following perspectives to build the radar domain dictionary:

(1) Radar domain terminology collection: special terms related to radar systems, equipment, software, etc.
(2) Radar domain discontinued word list: on the basis of the public discontinued word list, the auxiliary words that do not need to be considered in the radar test domain are removed, for example, debugging assistant, software, personnel, shall not, data, etc.
(3) Radar domain synonym collection: a collection of synonyms for radar terminology.
(4) Collection of typical radar software abnormal patterns: for example, target loss, track point overflow, array number and timing out of sync, etc.
(5) Radar software requirement features collection: for example, sector setting function, echo reception task, range, and speed measurement mode, etc.

There is an example of the domain dictionary for radar defects shown in the following Table 1.

Based on the radar domain dictionary and the inverse maximum matching algorithm, the software defective text word separation algorithm implemented in this paper is shown in Fig. 3:

Step 1: Assume that $X1$ is the radar software defect data string to be divided into words, the output string $X2$ is the empty set, and $\mathrm{Max}_{\mathrm{Word}}$ is the maximum word length in the radar domain dictionary.
Step2: If X_1 is the null set, then output X_2.
Step3: From the left of X_1, we pick out the candidate string X_{str} in sequence. The length of X_{str} is smaller than Max_{Word}.
Step 4: querying whether the string X_{str} belongs to the radar domain dictionary, if so go to step 5, otherwise go to step 6.Step 5: $X_2 = X_2 + X_{\mathrm{str}}$ + 'i', $X_1 = X_1 - X_{\mathrm{str}}$, and turn to step 2.
Step6: Eliminating the rightmost word of X_{str}.
Step7: If X_{str} is a single word, then turn to step5, otherwise, turn to step 4.

Table 1 dictionary for radar defects

No.	Domain terminology	Synonym of the domain terminology
1.	Silence zone	Silent area
2.	radio frequency	RF
3.	Main lobe	Main beam
4.	Fine tracking	Meticulous tracking
5.	……	……

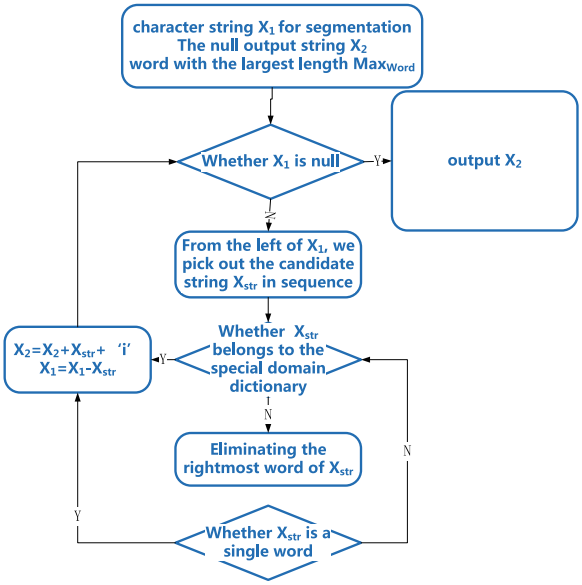

Fig. 3 Text segmentation approach based on the reverse maximum matching algorithm and radar domain dictionary

2.2 Modified LDA model and acquisition incorporating radar demand characteristics

2.2.1 The acquisition process of radar software defect data based on the LDA model

Based on Fig. 1, any document d (i.e., defect data) in corpus D (i.e., radar defect database) can be modeled with the help of the LDA model to generate topic probability distribution:

$$\overrightarrow{\theta}_d = \left(z_{d,1}, z_{d,2}, \Lambda, z_{d,n}\right) \tag{1}$$

Based on the above equation, the joint probability formula is obtained as follows:

$$p\left(\overrightarrow{w}_d, \overrightarrow{\theta}_d \middle| \alpha, \beta\right) = \prod_{n=1}^{N_d} p\left(w_{d,n} \middle| z_{d,n}, \beta\right) \bullet p\left(z_{d,n} \middle| \alpha\right) \tag{2}$$

The variables in Eq. 1 and 2 are explained as follows: First, assume that the radar defect database has D defect data (word segmentation set), and the total number of words is N_d, $w_{\mathrm{d,n}}$, and n represents the nth word in the dth data, each. The defect data consists of a mixture of k topics, then the topic-word probability distribution ϕ_k under each topic obeys the Dirichlet distribution with β as the parameter; θ_d is the document-topic distribution, and each defect data corresponds to a different topic distribution, and θ_d obeys Dirichlet distribution with α as the parameter, $Z_{\mathrm{d,n}}$ represents the specified distribution between topics and words within the defect data d, and $Z_{\mathrm{d,n}}$ obeys the polynomial distribution with θ_z as the parameter. According to Eq. 1 and Eq. 2, and the corresponding variables, based on the acquisition process of the radar software defects of the LDA model is described as follows:

1) For each defect data $d \in D$, according to $\theta_d \sim Dir(\alpha)$ (that is, θ_d obeys the Dirichlet distribution with α as the parameter), the polynomial distribution parameter θ_d is obtained;

2) For each topic $z \in k$, according to $\theta_d \sim Dir(\alpha)$, get the polynomial distribution parameter θ_d;
3) For the ith word $W_{d,i}$ in the defect data d:
   - According to the polynomial distribution $Z_{d,i} \sim \text{Mult}(\theta_d)$, get the topic $Z_{d,i}$;
   - According to the polynomial distribution $W_{d,i} \sim \text{Mult}(\theta_z)$, obtain the word $W_{d,i}$.

2.2.2 Modified LDA topic model incorporating radar demand characteristics

LDA uses a bag-of-words model to represent the text features extracted after dimensionality reduction, and when it is used for data text classification, the features are represented as probability vectors of each document-topic, and the similarity of probability vectors of each document is compared by Bayesian inference algorithm. The traditional LDA model does not take into account the word weight information in the domain context, and the topic assignment is skewed toward the topic to which the high-frequency words belong. In many cases, important terms with a strong domain background, such as radar software requirements, may not appear as often or as often as they should, making it difficult to be the output of topic-keyword. Therefore, with reference to the Labeled-LDA method [13], this paper modifies the original bag-of-words model in the LDA model by combining the radar software requirements information, increasing the weight and text length of the words characterizing the radar software requirements, so as to form an improved LDA topic model incorporating the radar software requirements features, which is implemented as follows:

Step 1: For the original defect data set $D = \{d_1, d_2, \ldots, d_n\}$, find the keywords $V = \{v_1, v_2, \ldots v_s\} (V \in D)$ that characterize the requirements of radar software, such as function name, interface name, interface type, and state name. At the same time, enhance the word frequency weights of these keyword terms.
Step 2: For each word v_i $(i = 1 \sim s)$ of the radar software requirement keyword word $v_i (i = 1 \sim s)$, match the relevant required information for it. For example, for function name v_i, text information such as function logic, interface name, and status name can be automatically added to the original defect dataset to form the expanded dataset $V' = \{v_1, v_2, \ldots, v_s, v_{s+1}, v_{s+2}, \ldots, V_S\}$. The final extended defect dataset $D' = D \cup V'$ (the length of the defect dataset is $S + n$).
Step 3: For the expanded radar software defect data set D', construct the probability vector model as shown in Eq. 3:

$$\begin{pmatrix} \theta_1 \times (\phi_{11}, \ldots, \phi_{1,\ S+n}), \ldots, \theta_z \times (\phi_{z1}, \ldots, \phi_{z,S+n}), \\ \theta_n \times (\phi_{S1}, \ldots, \phi_{S,\ S+n}) \end{pmatrix} \tag{3}$$

Among them, θ_1, θ_2, …, θ_n is the value of the n-dimensional document-topic probability vector $\overrightarrow{\theta}$, and ϕ_{z1}, ϕ_{z2}, …, ϕ_{zS} represents the probability distribution of the S core words in the term distribution corresponding to the zth topic. Therefore, the constructed probability vector model is $n \times S$-dimensional. This expanded defect data set retains the dimensionality reduction and noise reduction effect of the LDA model and

also incorporates the radar software requirement information, which has better semantic interpretability.

In order to obtain the word probability distribution in the above model, it is necessary to estimate the hidden parameters θ and ϕ. This paper uses the Gibbs sampling method and combines the expanded defect data set to obtain the estimated values of θ and ϕ. The estimation process of θ and ϕ can be regarded as the inverse process of the generation of defect data text, that is, in the case of a given defect data set, the estimated value of the hidden parameter is obtained through parameter estimation. The details are as follows:

Gibbs sampling is to determine the topic of each word. The hidden topic parameters can be obtained by counting topic frequency. This article assumes that the current word topic $\overrightarrow{z_{-i}}$ allocation is excluded, and the probability of the current word allocation to each topic is estimated according to the topic allocation of other words. The calculation formula is shown in Eq. 4:

$$p(z_i = k|\overrightarrow{w}, \overrightarrow{z}_{-i}) = \frac{p(\overrightarrow{w}, \overrightarrow{z})}{p(\overrightarrow{w}, \overrightarrow{z}_{-i})} \propto \frac{n^{(t)}_{k,-i} + \beta_t}{\sum_{t=1}^{V}(n^{(t)}_{k,-i} + \beta_t)} (n^{(t)}_{m,-i} + \alpha_t) \quad (4)$$

Among them, $z_i = k$ represents the topic k determined by the word i; $-i$ represents the set of other topics that do not include the word i; $n^{(t)}_{k,-i}$ represents the number of times the word t appears in the k topic; and $n^{(t)}_{m,-i}$ represents the number of times the defect data m appears on the topic k. Assume that each word If the theme is determined, then the estimates of $\theta$and ϕ can be carried out according to the Eq. 5:

$$\theta_{m,k} = \frac{n^{(k)}_{m,-i} + \alpha_k}{\sum_{k=1}^{K}(n^{(k)}_{m,-i} + \alpha_k)}, \phi_{k,t} = \frac{n^{(t)}_{k,-i} + \beta_t}{\sum_{t=1}^{V}(n^{(t)}_{k,-i} + \beta_t)} \quad (5)$$

Among them, $\theta_{m,\ k}$ represents the probability of topic k in the defect data m, $\phi_{k,\ t}$ represents the probability of the word t in the topic k, and iteratively calculates the topic distribution θ and topic word distribution ϕ of the defect data m.

3 Radar defect data topic acquisition and classification based on improved LDA model

This section first identifies the topic models of radar defect data based on the improved LDA model proposed in Chapter 2. On this basis, the similarity factor between each defect data and the topic model is calculated. Then, based on the similarity factor, the classification process of the radar software defect data is realized.

3.1 Defect data recognition based on modified LDA model

First, based on the improved LDA model in Chapter 2, the radar software defect dataset D is trained to learn and identify multiple topic models of software defect data. The details are as follows:

Input:

- Radar historical defect dataset D, containing N defect data
- Improved LDA model incorporating radar requirement features
- The expected number of implied topics k, the number of topic keywords s
- Radar domain dictionary: including radar domain terminology set, discontinued words list, and synonym set

Process:

1) Text segmentation of radar historical defect data set

According to the defect data text segmentation method based on radar demand features proposed in Section 2.1, with the help of radar domain dictionary, each defect data $d_i(i = 1 \sim N)$ in the radar historical defect data set D is subjected to text segmentation to obtain N a set of text segmentation vectors of defect data $D_S = \{d_{s1},d_{s2},\ldots,d_{sN}\}$.

2) The number of expected hidden topics k is determined

When applying the LDA topic model, it is technically difficult to determine the expected number of hidden topics k value. At present, there is no unified and standardized solution, and methods such as expert experience, experimental parameter adjustment, and optimal density model are mostly used. If the value of k is too small, the range of this paper is too broad; if the value of k is too large, meaningless garbage themes may be generated.

Based on the application background in the domain of radar software testing, this paper proposes a method for determining the k value of the LDA topic model based on radar demand characteristics, that is, firstly, according to the number of functions in the radar software project, the number of test types, the number of interfaces, and other demand information, the preliminary determination is made the approximate range of the number of topics k. On the basis of this range, the most appropriate k value is continuously adjusted and determined with the help of the variation of accuracy and recall values applied by the LDA topic model on defect data classification or defect prediction. The domain radar software has similarities. The k value can be determined in advance according to each radar domain as a priori knowledge of the RadarDCP method for new projects. Specifically, for a specific domain, the steps for determining the k value are shown as the following figure:

According to the above Fig. 4, we can determine the k value as the following steps:

- According to the number of functional items, the number of test types, the number of interface items, and other information in the radar software projects in this field, the radar experts and software developers jointly determine the initial value of the number of topics k (greater than 1);
- With the help of this initial value, apply the modified LDA topic model to radar defect data classification tasks or defect prediction tasks;
- The historical radar software projects in the field are selected, and the accuracy and recall rate of the LDA topic model are calculated for the radar defect data classification or defect task application, and the k value is continuously adjusted

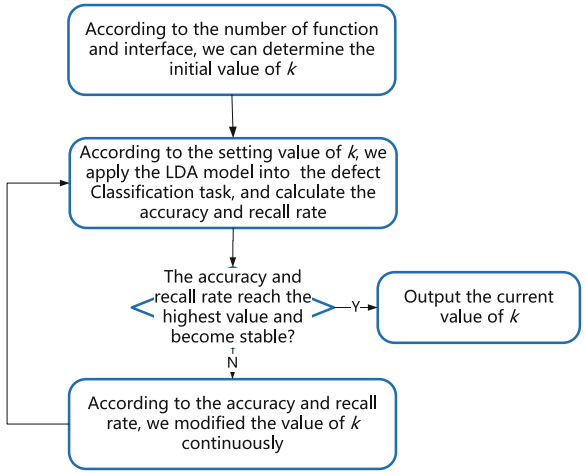

Fig. 4 The steps for determining the *k* value

based on the accuracy and recall rate until the classification accuracy and recall rate reach the highest value and become stable, at which time the current *k* value is considered to be the optimal value for the radar project in the current field.

- If the new software project belongs to the same radar system, the optimal *k* value of similar historical projects can be selected as the initial value of *k* for the new project, and then the optimal *k* value can be determined iteratively according to the above steps.

3) Topic identification of radar software defect data

The text segmentation vector set $D_S = \{d_{s1}, d_{s2}, \ldots, d_{sN}\}$ of the defect data, the expected number of hidden topics *k* value, is brought into the modified LDA model integrated with radar demand characteristics for training and learning, and *k* topic models are obtained, denoted as $D_M = \{D_{M1}, D_{M2}, \ldots\ldots, D_{Mk}\}$. Each topic model $D_{Mj}(j = 1 \sim k)$ is composed of *s* keywords, that is, the keyword vector can be characterized as: $D_{Mj} = \{d_{mj1}, \ldots, d_{mjk}\}$, where $d_{mjk} \in D_S$.

Output:

- *k* topic model D_M of radar software defect data

Keyword vector collection of topic model D_{Mj} $\{d_{mj1}, \ldots, d_{mjk}\}$

3.2 Defect data classification method based on topic similarity factor

Based on the *k* topic models of the generated radar software defect data, calculate the topic similarity between each defect data text word segmentation vector $d_{si}(i = 1 \sim N)$ and each topic model $D_{Mj}(j = 1 \sim k)$ Degree factor. This article intends to use JS scatter to calculate the similarity factor between the defective data and the

topic model with the help of Gibbs sampling method in Section 2.2.2. The Gibbs sampling method can estimate the topic probability of any defect data d_{si} in the defect data set D The distribution vector uses JS divergence and KL divergence [14] to calculate the topic similarity of the defect data. Therefore, the calculation formula of the similarity factor LS for the radar software defect data and the topic model is shown in Eq. 6:

$$LS = D_{JS}\left(\overrightarrow{\theta}_p, \overrightarrow{\theta}_q\right) = \frac{1}{2}\left(D_{KL}\left(\overrightarrow{\theta}_p, \frac{\overrightarrow{\theta}_p + \overrightarrow{\theta}_q}{2}\right) + D_{KL}\left(\overrightarrow{\theta}_q, \frac{\overrightarrow{\theta}_p + \overrightarrow{\theta}_q}{2}\right)\right) \tag{6}$$

Among them, D_{KL} represents KL divergence, and the calculation formula is shown in Eq. 7:

$$D_{KL}\left(\overrightarrow{\theta}_p, \overrightarrow{\theta}_q\right) = \sum_{k=1}^{K} \theta_{p,k} \ln \frac{\theta_{p,k}}{\theta_{q,k}} \tag{7}$$

Among them, p can represent the word segmentation vector of the defect data, and q represents the keyword vector of the topic model, $\overrightarrow{\theta}_p$, $\overrightarrow{\theta}_q$ is the topic distribution vector corresponding to the two respectively, and $\theta_{p,k}$ is the probability value of the defect data p belonging to the topic k.

The defect data classification process based on the topic similarity factor is shown in the Fig. 5.

According to Fig. 5, we illustrate the defect data classification process based on the topic similarity factor as the following steps:

For each defect data d_{si}, we calculate the similarity factor LS between it and each topic model D_{Mj}. If the value of LS is greater than the specified threshold, it can be determined that the defect data d_{si} has a high probability of belonging to the

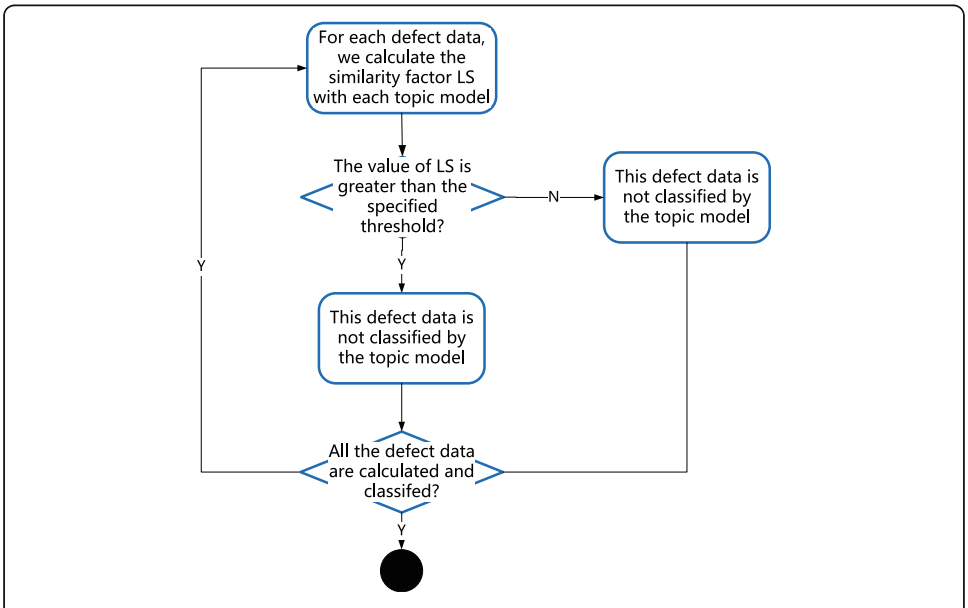

Fig. 5 The defect data classification process based on the topic similarity factor

topic model D_{Mj}, that is, the defect data d_{si} can be classified by the subordinate topic model. Repeat the above similarity factor calculation and analysis process, and for each defect data d_{si}, the corresponding subordinate topic model D_{Mj} can be found, which can also be called the calssfication model of the defect data d_{si}.

Finally, if m ($m \le k$) topic models have their own defect data sets, the radar software defect data can be classified, that is, the radar software defect data D is divided into m categories, corresponding to m topic models.

4 Typical case study

4.1 Experiment result of text segmentation based on the defect data text of the radar domain dictionary

4.1.1 (1) Defect data of radar software for test

1) A certain type of radar display and control software test questionnaire: a total of 324 items;
2) The problem description text in the test question sheet is used as the training test object, and the defect data has been standardized to form 324 items of radar defect data;
3) Organize technical staff in the radar domain to pre-segment and label the 324 items of problem description texts, forming the standard radar defect text segmentation set.

4.1.2 (2) Text segmentation technology for comparison

The common techniques for Chinese text segmentation include conditional random domain model, hidden Markov model, and maximum entropy model. These three techniques have their own advantages and disadvantages, but the conditional random domain model is the most researched technique in recent years. Therefore, this paper selects the conditional random domain model as the Chinese text segmentation technique for comparison.

4.1.3 (3) Test index for text segmentation comparison

The following three main indicators are used to evaluate the performance of the Chinese text segmentation system:

1) Word segmentation accuracy rate P = (the number of text segmentation accurately achieved by the algorithm)/(the number of all text segmentation achieved by the algorithm) × 100%.
2) Word segmentation recall rate R = (the number of text segmentation accurately achieved by the algorithm)/(the number of text segmentation in the standard radar defect text segmentation set) × 100%.
3) Comprehensive index $F = 2PR/(P + R)$
4) The definitions of the above indicators in the 6.2 and 6.3 tests are similar to those of 6.1, so the introduction will not be repeated.

4.1.4 (4) Test process of text segmentation

Step 1: Text segmentation method training

Firstly, the general domain annotated corpus (namely People's Daily annotated corpus (http://www.icl.pku.edu.cn/icl_res/)) provided by the Institute of Computer Linguistics of Peking University and 200 items of radar defect data texts were selected as the training set. The text segmentation method CRF-SE based on conditional random domain model and the text segmentation method RadarDCP-SE based on the radar domain dictionary and the reverse maximum matching algorithm proposed in this article are studied, and the trained CRF-SE and RadarDCP-SE are obtained.

Step 2: Text segmentation based on the test set

Then, 100 items from the Radar Defects Data Text were selected as the test set, and two text segmentation methods, CRF-SE and RadarDCP-SE, were applied to each test data respectively, and perform text segmentation and remove the stop words to obtain the radar defects text segmentation results.

At the same time, 100 items of data are selected as the test set from the general domain annotated corpus, and the two-word segmentation methods CRF-SE and RadarDCP-SE are respectively applied to each test data. After text segmentation is performed and the stop words are removed, obtaining the universal domain text segmentation result.

Choose a radar defect data as an example to illustrate that the text segmentation results under the two methods of CRF-SE and RadarDCP-SE are as follows:

The original text of one item of Radar Defect Data is as follows: a select 4-degree target for clutter zone setting and the fixed point setting is consistent with the clutter zone setting. If the elevation angle range of the clutter zone is set to a normal value, the target in the silent zone can be successfully seen. Otherwise, the silent zone target cannot be seen.

- The CRF-SE segmentation results for the above original text are as follows: "Clutter," "Zone," "Selection," "Target," "Fixed point," "Clutter," "Zone," "Consistent," "Clutter," "Zone," "Elevation angle," "Range," "Normal value," "Silent," "Zone," "Target," "Silent," "Zone," "Target"
- RadarDCP-SE segmentation results for the above original text are as follows: "Clutter Zone," "Select," "Target," "Fixed Point," "Clutter Zone," "Consistent," "Clutter Zone," "Elevation angle," "Range," "Normal Value," "See," "Silent Zone," "Target," "Silent," "Zone," "Target."
- Step 3: Comparison of word segmentation test results

According to the general domain text segmentation results and the radar defect text segmentation results obtained by the two methods of CRF-SE and RadarDCP-SE, three indicators of P, R, and F are calculated, as shown in Tables 2 and 3:

4.1.5 (5) Experimental conclusion and analysis

From Tables 2 and 3, the following experimental conclusions can be drawn:

Table 2 Comparison results between CRF-SE and RadarDCP-SE on texts of radar defects

Word segmentation method	Accuracy rate *P*	Recall rate *R*	Comprehensive index *F*
CRF-SE	0.7789	0.7622	0.7704
RadarDCP-SE	0.8976	0.8761	0.8867

In the radar defect text, the RadarDCP-SE method has significantly improved over the CRF-SE method in terms of accuracy *P*, recall *R*, and comprehensive index *F*. In other words, the RadarDCP-SE method is better than the CRF-SE method in all three indexes.

The analysis of the experimental results is as follows:

As can be seen from the examples in step 2, the CRF-SE method does not work well for the segmentation of the terminology of the radar domain. For example, theclutter zone is divided into the clutter and zone, and the silent zone is divided into "silence" and "zone," while the RadarDCP-SE method can well realize the word segmentation of proper nouns in the radar field.

In addition, the results of CRF-SE and RadarDCP-SE methods are almost similar for generic domain texts, which indicates that the Radar domain dictionary used in the RadarDCP-SE method is not well supported if the domain constraint is removed, but at least this shows that the RadarDCP-SE method is not worse than the CRF-SE method, and it also shows that the applicability of the method proposed in this article can also meet the needs of text segmentation in the general domain.

4.2 Radar defect data classification experiment based on modified LDA topic model

4.2.1 (1) Defect data of radar software for test

1) Test question list of a certain type of radar display and control software: a total of 90 items;
2) In advance, the test questionnaires are manually classified into 9 categories according to the functions they belong to, forming a standard classification set, that is, the 90 test questionnaires belong to 9 functions on average.

4.2.2 (2) Classification method for comparison

Data classification methods for comparison include the following: traditional LDA topic model (denoted as LDA-CF), modified LDA topic model (denoted as RadarDCP-CF) that incorporates radar demand features proposed in this paper, and mainstream classification algorithm support vector machine SVM (marked as SVM-CF).

4.2.3 (3) Model parameter setting

This article uses the python language to implement the LDA topic model under the Windows7 operating system. The main parameter settings for the LDA topic model are

Table 3 Comparison results between CRF-SEand RadarDCP-SE on texts of common domains

Word segmentation method	Accuracy rate *P*	Recall rate *R*	Comprehensive index F
CRF-SE	0.8133	0.8043	0.8088
RadarDCP-SE	0.8176	0.8086	0.8131

as follows: Gibbs sampling method is used for parameter estimation. The ratio of the training set and the test set of the radar software defect data is set to 5:1. The document-topic probability distribution parameter α is 0.1, the topic-word probability distribution parameter β is set to 0.01, and the number of keywords under each topic is set to 10.

4.2.4 (4) Classification test process

Step 1: Set the expected number of topics of the LDA topic model k

In the standard classification set, it is manually classified into 9 categories. Therefore, set the expected number of topics k of the two topic models of LDA-CF and RadarDCP-CF to 9.

Step 2: Segmentation of defect data text

With the help of the Defect Data Text Segmentation Based on Radar domain Dictionary algorithm proposed in this paper, 90 defect data are segmented.

Step 3: Classification process based on model training

The three classification models of LDA-CF, RadarDCP-CF, and SVM-CF are trained and learned on the word segmentation set of 90 defect data to form the corresponding classification results.

Among them, taking the RadarDCP-CF model as an example, the classification results of the defect data are as follows:

Among the 90 items of defect data, there are 6 items of data (as shown in the Table 4, the original data text is longer and simplified to a certain extent) related to the initialization function. The RadarDCP-CF model groups these 6 items of data (seen in Table 4) into model S for the same topic is as follows:

Topic model S: ("0.046'Receive" + 0.035 × "Power-on initialization" + 0.035 × "Station position" + 0.029 × "Configuration file" + 0.028 × "Identification" + 0.025 × "Radar A" + 0.025 × "Radar B"+ 0.024 × "target" + 0.024 × "read" + 0.022 × "process").

Table 4 Six radar software defects of initialization

Serial number	Radar software defect data text
1.	When powering on, the "Station ID" read by the station software from the configuration file is an undefined value
2.	During initialization, the modes read from the configuration file by each station are inconsistent
3.	When powering on, the 'radar working mode' received by each station is inconsistent
4.	The value of 'initial working mode' received by the software is an undefined value
5.	During the power-on initialization process, the software of the two stations read the same "station ID" from the configuration file, that is, the same radar A or radar B
6.	During the power-on initialization process, multiple program instances are run at the same time on one station

Table 5 Comparison results between LDA-CF, RadarDCP-CF, and SVM-CF on texts of radar defects

Classification	Accuracy rate *P*	Recall rate *R*	Comprehensive index *F*
LDA-CF	0.6202	0.7386	0.6742
RadarDCP-CF	**0.8019**	**0.8743**	**0.8365**
SVM-CF	0.7287	0.6938	0.7108

According to the analysis of the topic model *S*, it can be seen that the keyword set contained in it is relatively close to the description of the initialization function, such as "power-on initialization," "receive," and "configuration file." Therefore, these 6 items of defect data are initialized. It is classified as topic model S, which is in line with our expectations. In the standard classification set, these 6 defect data are also classified into the same category, which belongs to the initialization function.

Step 4: Comparison of classification test results

According to the three classification algorithms of LDA-CF, RadarDCP-CF, and SVM-CF, the classification test is carried out on 90 defect data, and the calculation of *P*, *R*, and *F* indicators is performed, as shown in Table 5.

4.2.5 (5) Experimental conclusion and analysis

From Table 5, the following experimental conclusions can be drawn: in the classification of radar defect data, the RadarDCP-CF algorithm has a significant improvement in accuracy rate *P*, recall rate *R*, and comprehensive index *F* compared to the other two algorithms. Among the three indicators, the RadarDCP-CF algorithm the classification performance is the best.

In step 3, the classification result of the RadarDCP-CF algorithm shows that it is consistent with the standard classification set. In contrast, the classification result of the LDA-CF algorithm classifies the fourth data item into other models, but not into the valid classification result. This is because the words "power on, initialize" do not appear in the fourth data item. However, the interface data of "start mode" belongs to the function of "power-on initialization." In the RadarDCP-CF algorithm, "operating mode" and "power-on initialization" is automatically associated together, and the weight of "operating mode" in the classification is increased, so the accurate classification results are obtained.

5 Conclusion

This paper proposes a new method for classifying radar software defects based on an improved LDA topic model, including the following: firstly, we propose the radar domain dictionary to achieve accurate classification of radar software defects; then, we propose the modified LDA topic model incorporating radar requirement features to obtain potential topics oriented to requirements such as function names and interface names and improve the topic acquisition accuracy of the LDA model. Then, based on the obtained topic model, the classification of radar software defect data is realized; finally, the research of engineering application is carried out for typical radar software defect data. Compared with the conditional random domain model-based classification

method, the accuracy and recall rate of the radar domain dictionary-based classification technique proposed in this paper are improved by 15%; compared with the traditional LDA topic model-based data classification method, the improved LDA topic model-based radar defect data classification method proposed in this paper is improved by 24% in terms of accuracy and recall rate.

The application research shows that the proposed software defect classification method is better than the existing methods in the application of radar domain defect data, which can well conform to the complex domain characteristics and failure laws of radar software defect data, realize the accurate classification and prediction of radar software defect data, and then improve the efficiency and quality of radar software testing and design work.

Authors' contributions

XL and YY carried out the main idea and algorithm of the intelligent radra software defec classification approach based on the latent Dirichlet allocation topic model. HL and JC participated in the Modified LDA topic model incorporating radar demand characteristics and the typical case study. CL participated in the Radar defect data topic acquisition and classification based on improved LDA model. SW and RY participated in the typical case study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Nanjing Research Institute of Electronics Technology, Nanjing 210013, China. [2]School of Software, Beihang University, Beijing 100191, China. [3] School of Reliability and Systems Engineering Beihang University Beijing 100191 China .

Received: 23 March 2021 Accepted: 2 July 2021
Published online: 20 July 2021

References

1. L.N. Gong, S.J. Jiang, L. Jiang, Research progress of software defect prediction. Ruan Jian Xue Bao J. Software **30**(10), 3090–3114 (2019) (in Chinese). http://www.jos.org.cn/1000-9825/5790.htm
2. T. Wang, *Test case reuse based on software requirement* (University of Chemical Technology, Beijing, 2017) (in Chinese)
3. Z.G. Zhang, B.L. Xu, X.H. Qin, Reuse-oriented modeling of test cases for spaceflight TT&C software. J. Spacecraft TT&C Technol. **6**, 46–50 (2011) (in Chinese)
4. L.Z. Meng, H. Wang, Y.Z. Xue, et al., Research on automatic generation of software failure modes. J. Front. Comput. Sci. Technol **12**(11), 63–71 (2018) (in Chinese)
5. L. Xi, Z. Zhiyong, L. Haifeng, L. Chang, S. Wang, Defect prediction of radar system software based on bug repositories and behavior models. Int. J. Perform. Eng. **16**(2), 284–296 (2020)
6. Liu X, Liu C, Li HF, Liu Y, Wang SL. Defect prediction of radar system software based on system-theoretic accident modeling process. Modern Radar, **42**(11), 87–92 (2020) (in Chinese)
7. Y.S. Chen, Y.T. Shi, Domain specific Chinese word segmentation. Comp. Eng. Appl. **54**(17), 30–34 (2018) (in Chinese)
8. X. Feng, Comparison of methods for integrating lexicon information in Chinese word segmentation. Appl. Res. Comput. **36**(1), 8–10 (2019) (in Chinese)
9. Shi Y. *Research on Chinese word segmentation based on deep learning*. (Nanjing University of Posts and Telecommunications, Nanjing, 2019)(in Chinese)
10. G. Xu, H.F. Wang, The development of topic models in natural language processing. Chinese J. Comput. **34**(8), 1423–1436 (2011) (in Chinese)
11. N. Liu, Y. Lu, X.J. Tang, et al., Multi-document summarization algorithm based on significance topic of LDA. J. Front. Comput. Sci. Technol **9**(2), 242–248 (2015) (in Chinese)
12. H.J. Luo, X.H. Ke, Automated scoring Chinese subjective responses based on improved-LDA. Comput. Sci. **44**(z11), 102–105 (2017) (in Chinese)
13. W.B. Li, L. Sun, D.K. Zhang, Text classification based on labeled-LDA model. Chinese J. Comput. **31**(4), 620–627 (2008) (in Chinese)
14. W. Hu, Y.H. Jing, Recommendation algorithm based on fusion of KL divergence and JS divergence similarity. J. Harbin Univ. Commerce **36**(1), 48–53 (2020) (in Chinese)

Publisher's Note