M. V. TKACHUK, R. A. GAMZAEV, I. O. MARTINKUS, S. D. IANUSHKEVYCH

METHODS AND TOOLS FOR DYNAMIC REQUIREMENTS CATALOG MANAGEMENT IN AGILE SOFTWARE DEVELOPMENT

A method for managing dynamic requirements catalog in agile software development, especially on example of Scrum-methodology is proposed. Popular approaches to solving this problem are reviewed. The proposed approach is based on the combined usage of the latent semantic analysis and analytical hierarchy process, it allows to evaluate the given textual software specification with respect to their possible redundancy and possible logical conflicts. Besides that this approach supports the decision making procedure to prioritize the requirements taking into account their functionality importance for target software product. The effectiveness of the proposed method was tested on the test case.

Keywords: Agile, latent semantic analysis, dynamic catalog of requirements analytic hierarchy process.

Introduction: Problem Actuality and Research Goal. Requirements management (RM) is one of the most important and weak-formalized disciplines in the modern software engineering (SE), because of permanent changes in an application business logic and/or in a projects configuration. This is the real challenge for any new software system to be designed “from scratch” or for already existing software systems. To resolve these problems the several adaptive SE-methodologies called as agile software development (ASD) are proposed [1]. The main concepts of any ASD-methodology supposed to meet requirement changes in a following way:
(1) organization of a permanent and closed collaboration between several actors involved in ASD: stakeholders, analysts, developers, end-users, etc.;
(2) usage of an iterative approach for the project development;
(3) providing a requirements traceability in order to reflect all requirement changes into appropriate projects artifacts.

In all ASD-methodologies, taking into account their base principles (1)–(3) is supposed, that the management of permanent changed in software requirement specifications (SRS) should be provided in an adaptive mode. It means necessity in appropriate way quickly to change the design solutions, development environment, the organization forms of a developer team, etc. Such dynamic and multidimensional vision for Requirements management is well shown on qualitative model of this process proposed by S. Ambler [2], which is shown below in Fig. 1.

Figure 1 – Conceptual scheme for adaptive requirements management by S. Ambler [2]

It is necessary to note that besides some empirical recommendations concerning the problem how to build such a dynamic requirements catalog (DRC), in recent literature there are no more or less proved suggestions about possible quantitative effectiveness estimations for this purpose.

This problem is actual also in a broader sense, because DRC is a high-level requirements representation for any domain, independently from their specification method with any notation (UML, IDEF0 etc.). So DRC is essentially an example of such important concept in Agile methods as initial requirements[3], which are required for building on their base Domain Model, which in turn is basis for such modern and effective software development approach as Domain Driven Development (DDD) [4]. But in [3] don’t proposed any quantitative methods and evaluations as for priority determination of particular requirements, as well as solving of such important tasks, as the elimination of logical contradictions and data duplications in text specifications.

That is why the main research aim of this paper is to consider some specific features of requirements management in ASD, to analyze available formal methods for requirements evaluation and prioritization, and to
propose the complex of algorithms and software tools, which can be used for DRC processing.

Specific features of requirements management in Agile-software development on Scrum-methodology example. As already mentioned above, ASD methodologies are widely used due to new nature of software projects and due to permanent changes in the SRS. ASD implies refusal from a huge of project documentation, but really needs support for automated of stakeholders knowledge processing, in order to reduce the time for adaptation to requirements changes and to minimize risks related to them.

Exactly because of these reasons in the modern SE-concepts process control methods called also as “software cybernetics” (e.g. in [5]) are used. On Fig. 2 the cybernetics-centered scheme is shown [6], which represent one of the most used ASD-methodology, namely Scrum method [1, 2]. There are 2 feedback loops included in this control scheme: (1) daily process control loop, which is organized basing on sprint backlog (SB) and using some source code quality metrics, (2) iteration process control loop, which is provided basing on product backlog (PB) and using some requirements quality metrics. The SB and PB both collect the selected requirements to be met in final software product. In the control loop (1) usually some IDE, e.g. the Eclipse is used by developers team, while in the control loop (2) an appropriate requirements management system (RMS) can be exploited by stakeholders (or Product owner – in Scrum project) and by domain analysts.

In [6] the approach to creation so-called advanced traceability matrix is elaborated, which allows to take into account all developers activities in time-oriented data model, and the appropriate CASE-tool designed to combine the features of typical RMS and the functionality of integrated development environments (e.g. Eclipse). But the problem how to form and to maintenance an appropriate DRC is still open, and below we propose methods and software tools for this problem solution.

Overview of the text specifications analysis and prioritization methods for the composition of the dynamic catalog of requirements. Current approaches to the treatment of software requirements formulated in natural language can be divided into 3 main groups:

1. The construction of formal models for the original text descriptions of requirements, using the algebraic notation or the apparatus of predicate logic ([7]), for subsequent qualitative analysis and automatic verification;
2. Development and application of different methods for building automated visual, structural and graphical forms of presentation of software requirements: e.g. in the form of ER-diagrams and UML-diagrams, based on their initial descriptions of natural language etc., this group includes many works such as [8];
3. Various methods of logical and linguistic analysis of the semantic features of natural language texts of specifications that are used to extract hidden knowledge, establish semantic dependencies, eliminate logical contradictions, etc. [9].

Obviously, considering the properties of agile software development in general, and the scheme of the Scrum-project in particular, out of the above methods for processing requirements text descriptions, in the context of this work the methods of the third group should be considered, as no formal requirements specification (covered by the Group 1 methods), or their diagrammatic documentation (covered by Group 2 methods) constitute essential prerequisites (conditions) for the successful implementation of agile-projects [10–11].

In the selected group of methods for processing text requirements, taking into account the specific tasks in the control loop of the process of forming the dynamic catalog of requirements DRC, for the solution of the first of them – namely, for eliminating possible logical inconsistencies and redundancy (duplication) of data – different data mining methods [12], the methods of the analysis of hypertext links [13], and the various modifications of the method of latent semantic analysis (LSA) [14] can be effectively applied.
The LSA method is a statistical method for processing natural language texts, the purpose of which is to establish the relationships between documents that are the part of a collection (case), and a set of some terms contained in these documents. It is possible to identify the thematic proximity of certain terms, which can then be used to calculate statistical estimates of the thematic proximity of the entire documents. As the input data the LSA method uses so-called "term-to-documents" matrix ("term" refers to a specific lexical unit), which reflects the frequency of occurrence of the particular term in the text of the particular document. Each column of the matrix refers to the certain document, and each line – to the certain term. Various approaches (metrics) are used for defining the value of a certain matrix element on the intersection of the $i$-th row and $j$-th column. One of such metrics is so-called TF-IDF (Term Frequency – Inverse Document Frequency) – the metrics which is used for the estimation of the importance of the term for the certain document from a collection [15]. To solve the second problem of the initial processing of the text description of requirements in the control loop of forming the dynamic catalog – namely, to prioritize the requirements, we first should take into account some methodological features of this process in the context of the agile methodologies of software development. It should be noted that in many works concerning the characteristics of the requirements engineering in agile methodologies it is directly stated that the creation and use of the appropriate mechanisms for prioritization and reprioritization of requirements is a key factor in the success of the projects making use of the agile methodologies. At the same time, in the agile projects (XP, Scrum, ASD), there are characteristic differences from the traditional methodologies (cascade, spiral, etc.) in the requirement prioritization process (RPP), these differences are analyzed in detail in [15], and are summarized in table 1.

<table>
<thead>
<tr>
<th>RPP characteristics</th>
<th>Traditional methodologies</th>
<th>Agile methodologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>When RPP is carried out</td>
<td>As a rule, once, after the analysis stage and before the implementation stage</td>
<td>Before each design iteration, or even during the iteration</td>
</tr>
<tr>
<td>Who initializes / leads RPP</td>
<td>Software developers, with the assistance of the manager of the project and other stakeholders</td>
<td>The software customer with the help of the manager of the agile development group (Scrum master)</td>
</tr>
<tr>
<td>Goals / prerequisites of carrying out RPP</td>
<td>The control for the process of carrying out the project</td>
<td>Support of the max. business significance of the results of software development / more precise specification of the amount of works of the current iteration of the project</td>
</tr>
</tbody>
</table>

The presence of all these factors leads to the fact that for RPP various expert methods are generally used, a fairly complete overview of which is given in [15]. Among them the following methods can be considered:

1. The method of the direct pairwise analysis – is carried out by pairwise comparing of separate requirements until the requirement with the highest priority appears at peak of their initiating list; owing to it this approach is applicable only in projects with a small amount of well-structured and plain text descriptions of requirements;

2. Wiegers matrix approach – is based on representing the set of software functional requirements in the form of the special table (matrix), the elements of which represent estimations, and are directly proportional to their technical significance for the final software products and inversely proportional to the cost and possible risks of their implementation; as a result of the analysis of such matrix it is possible to select the requirements with a maximum priority taking into account a compromise choice based on the values of the corresponding elements of this matrix;

3. The method of creation of a binary search tree – represents the approach based on a well-known method of binary search from the area of information search which can be adapted to the tasks of the prioritization of requirements, as a result of it the initial list of requirements can be represented in a way when the requirements with maximal and minimal priorities will be positioned in the ordered list.

Among the other methods it is important to mention the analytic hierarchy process [16]. This method will be used in this paper to prioritize DRC requirements after LSA preprocessing.

The complex procedure for constructing the dynamic requirements catalogue. The corresponding algorithms implementing its separate stages are given in more details below.

**LSA requirements processing algorithm.**

Step 1: Forming the matrix of the occurrence frequency for the $i$-th term in $j$-th requirements text:

$$
M[x_i] = \begin{bmatrix}
    x_{i1} & ... & x_{i1} & ... & x_{in} \\
    ... & ... & ... & ... & ... \\
    x_{i} & ... & x_{ij} & ... & x_{in} \\
    ... & ... & ... & ... & ... \\
    x_{n1} & ... & x_{nj} & ... & x_{nn}
\end{bmatrix},
$$

where $x_{ij}$ – is the occurrence frequency of the $i$-th term in the $j$-th requirement text.

Step 2: Removal of the columns in the matrix $M[x_{ij}]$ where the term is included in only one requirement, resulting in reduced matrix $M'[x_{ij}]$ with dimensions $m \times n$. 

\[m \times n\]
Step 3: Creating the reduced matrix $M'[x_i]$ by means of a method of singular expansion [11], i.e. to present a matrix in a type:

$$M'=UWV^T,$$

(2)

where $U$ and $V$ are the orthogonal matrixes.

Step 4. By selecting $k$ largest singular values $w_j$, the approximation of an initial matrix by the matrix of the smaller rank $k$ can be made. Performing the procedure of the rank reduction allows to eliminate surplus information, the rank value $k$ is defined heuristically [13]. Matrixes are displayed as follows:

$$W'=egin{bmatrix} w_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & w_j & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & w_1 \\ \end{bmatrix},$$

(3)

$$U'=[[u_1], \ldots, [u_k]],$$

(4)

$$V'=[[v_1], \ldots, [v_k]].$$

(5)

Step 5: Obtaining the final matrix $A$ with the frequencies of the occurrence of $i$-th term in $j$-th requirements description that contains only $k$ first linearly independent elements:

$$A=U'W'V'^T.$$ 

(6)

In contrast to the original matrix $M$, the resulting matrix $A$ is not sparse [13], which allows calculating dependencies between requirements documents even if the documents have no common terms.

Step 6: For finding dependencies between texts of different requirements on the basis of the obtained matrix $A$ it is necessary to calculate the Piersons correlation coefficient [17] between requirements documents:

$$R = [r_{ij}], \text{ for } \forall i > j, i, j = 1, n \text{ and } r_{ij} = 1 \text{ if } i = j.$$ 

As a result of the given step the triangular matrix $R$ containing correlaions coefficients is obtained:

$$R = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{nn} & \cdots & r_{nm} & \cdots & r_{nn-1} \end{bmatrix}. $$

(7)

The values of the correlation coefficients $r_{xy}$ between two documents $x$ and $y$ are calculated on the basis of the data received in 8. Denoting pairwise the compared columns of matrix $A$ as vectors $x = \{x_i\}$ and $y = \{y_i\}$, we receive:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^{2}} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^{2}}}, \quad r_{xy} \in [-1; 1]$$

(8)

where: $i = 1, n, j = 1, n, \bar{x}, \bar{y}$ – average value of samples $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Step 7: Determining the critical value of the Pearson correlation coefficient $r_{cr}$ for a given level of significance. Here the positive value of correlation coefficients between the requirement $x$ and $y$, $r_{xy} > r_{cr}$ which exceeds the positive critical value $r_{cr} > 0$, means the existence of the contradiction in the description of these requirements. The negative value of the correlation coefficients between the requirement $x$ and $y$, $r_{xy} < -r_{cr}$, means that these requirements have redundancy in their description, and one of the requirements can be removed.

As a result of the execution of this algorithm (possibly, in multiple iterations) it is possible to determine all requirements whose descriptions contain logical contradictions and overlapping texts.

**AHP-based requirements processing algorithm.**

1 stage: generate criteria which will be used for an estimation of the requirements importance.

2 stage: generate a matrix of pairwise comparison of the importance criteria, we denote it as $M[x_i]$ with dimensions $m \times m$, where $m$ is the number of criteria. As an estimation of the importance measure we will use score estimations. The resulting matrix by definition should possess the following properties:

All diagonal elements are equal to 1

All remaining elements possess the property of reverse symmetry, i.e. the values of matrix elements symmetric to the main diagonal are reverse values [16].

3 stage: to generate $m$ reverse-semantic matrixes with dimensions $n \times n$ where $m$ – the number of criteria, $n$ – the number of requirements. The matrixes’ values are estimations of pairwise importance of each requirement for each criterion:

$$A_k = [a_{ij}], \quad k = 1, m; \quad i, j = 1, n.$$ 

(9)

4 stage: For $M$ and $A_k$ matrixes obtained at the previous steps calculate local priorities of criteria by the formulas (10) and (11) respectively:

$$L_i = \left( \prod_{j=1}^{n} x_{ij} \right)^{\frac{1}{n}}, \quad i = 1, n$$

(10)

$$AL_i = \left( \prod_{j=1}^{n} a_{ij}^k \right)^{\frac{1}{n}}, \quad i = 1, n$$

(11)

where $x_{ij}$ – an element of matrix $M$;
\[ a_{ij}^k \text{ - elements of a matrix } A_i; \]
\[ \prod_{j=1}^{n} x_{ij}^k \text{ and } \prod_{j=1}^{n} a_{ij}^k \text{ - product of all elements in } i\text{-th line.} \]

5 stage: For \( L_i \) and \( AL_i^j \) values obtained at the 4th step carry out the normalization of the local priority criteria for all pairwise comparisons, for this purpose it is necessary to calculate local priorities of criteria by the formula (12):

\[ L_i = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}, \quad i = \overline{1, m}, \tag{12} \]
\[ AL_i^j = \frac{a_{ij}^k}{\sum_{j=1}^{m} a_{ij}^k}, \quad i = \overline{1, m}. \tag{13} \]

On the base of the normalized calculated values it is possible to form matrices \( L \) and \( AL \):

\[ L = L_{ij}, \quad i = \overline{1, m}, \tag{14} \]
\[ AL = AL_{ik}, \quad j = \overline{1, n}, \quad k = \overline{1, m}. \tag{15} \]

6 stage: calculate global priorities, as matrix products – the formula (16):

\[ G_j = AL \times L, \quad j = \overline{1, n}. \tag{16} \]

The matrix \( G_j = \{w_1, \ldots, w_n\} \) contains coefficients defining requirements priorities.

As a result, by means of AHP algorithm the weight coefficients defining the importance of the requirements based on the selected criteria have been obtained.

The instrumental complex for implementing the procedure of constructing the dynamic requirements catalogue and the test calculations results. As the frequency of solving the task of constructing and modifying the dynamic requirements catalogue is not real-time, but assumes the possibility of performing the analysis and the formation of such requirements set synchronously with the planning of the next iteration of the agile-project, i.e. in case of Scrum-methodology its execution frequency corresponds to the period of 2-4 weeks, for the implementation of the corresponding procedure it is possible to use already existing software tools, such as MS Excel from the MS Open Office [18] tool suite and Matlab math package [19].

The input data in this schema are unprocessed initial requirements to the developed software product. Requirements are received from the product owner who forms the list of requirements and highlights keywords for each requirement. The developed application allows to form an occurrence matrix of keywords which forms the basis for calculation of correlation coefficients between requirements. The MatLab environment is used for this purpose, the data is imported from Excel file into this environment.

For carrying out the numerical experiment the table of requirements and the highlighted keywords has been made, see table 2.

<table>
<thead>
<tr>
<th>№</th>
<th>The requirement (the initial text)</th>
<th>Keywords</th>
<th>Id.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>«The automated calculation system (ACC) should be expected to be used for processing the data about corporate customers, physical persons, legal bodies».</td>
<td>«The automated calculation system», «customers», «physical persons», «legal bodies».</td>
<td>r1</td>
</tr>
<tr>
<td>2</td>
<td>“The capability of working with groups of clients, supporting functions” friends “and” family “and the cost-based user groups that are grouped based on arbitrary attributes, adding / removing subscribers in a group as the operator and authorized person</td>
<td>“Customer”, “add/remove”, “subscribers”.</td>
<td>r2</td>
</tr>
<tr>
<td>3</td>
<td>«Creating complex tariff plans with the participation of services that belong to different personal accounts».</td>
<td>«Tariff plan”, “Personal account”.</td>
<td>r3</td>
</tr>
<tr>
<td>4</td>
<td>«Convenience of tariff plans creation, a visual programming environment».</td>
<td>«Tariff plan”, “Programming environment”.</td>
<td>r4</td>
</tr>
<tr>
<td>5</td>
<td>«Formation of the new, modification and removal of existing tariff plans by ACC operator».</td>
<td>«Tariff plan”, “modification”, “removal”.</td>
<td>r5</td>
</tr>
<tr>
<td>6</td>
<td>«Explicit representation of tariff plans in automated ACC workplace for different types of ACC users».</td>
<td>«Tariff plan”, “Types of users”.</td>
<td>r6</td>
</tr>
<tr>
<td>7</td>
<td>«Creation of tariff plans of arbitrary complexity (including in the TP an unlimited number of tariff zones)».</td>
<td>«Tariff plan”, “Tariff zone”.</td>
<td>r7</td>
</tr>
</tbody>
</table>

The calculation final results of Pierson mutual correlation are shown in table 3. For a significance level \( r_{cr} = 0.874 \) values that have the correlation above critical, are of interest for the further analysis. As correlation between \( r_1 \) and \( r_6 \), \( r_3 \) and \( r_4 \), \( r_3 \) and \( r_7 \) is above critical value these requirements have been analyzed and the conclusion was made about the data redundancy in their description. So the requirements \( r_4 \), \( r_6 \) and \( r_7 \) have been excluded from the log of requirements. After this, for each criterion, the pairwise comparison of the requirements importance has been performed, their local priorities, and the normalized priorities have been calculated, and then the generalized global priority has been calculated. The results are shown in table 4.
As a result of applying LSA and AHP methods with a complex procedure proposed earlier (fig. 2 in the previous section of this paper) the dynamic catalogue of requirements is formed, the requirements are then transferred then into adaptive tracing control loop.

**Conclusions and Future Work.** In this paper some existing approaches to requirements assessment and prioritization were analyzed, and the appropriate methods and software tools were developed to management dynamic requirements catalog (DRC) management in agile software development, especially on example of Scrum-methodology. The proposed approach is based on the combined usage of the latent semantic analysis and analytical hierarchy process, was allows to evaluate the given textual software specification with respect to their possible redundancy and possible logical conflicts. Besides that this approach supports the decision making procedure to prioritize the requirements taking into account their functionality importance for target software product. Our future work concerns final implementation of the proposed CASE-tool, as well as improving the algorithms for DRC management using other experts estimation methods in order to make its more precise and efficient.

### Table 3 – Pierson correlation

<table>
<thead>
<tr>
<th></th>
<th>r1</th>
<th>r2</th>
<th>r3</th>
<th>r4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>1</td>
<td>-0.423</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>1</td>
<td>0.362</td>
<td>0.430</td>
<td>1</td>
</tr>
<tr>
<td>r4</td>
<td>1</td>
<td>0.362</td>
<td>0.897</td>
<td>1</td>
</tr>
<tr>
<td>r5</td>
<td>1</td>
<td>0.507</td>
<td>0.557</td>
<td>0.640</td>
</tr>
<tr>
<td>r6</td>
<td>1</td>
<td>0.936</td>
<td>0.485</td>
<td>0.485</td>
</tr>
<tr>
<td>r7</td>
<td>1</td>
<td>0.362</td>
<td>0.430</td>
<td>0.919</td>
</tr>
</tbody>
</table>

### Table 4 - Results of LSA application

<table>
<thead>
<tr>
<th>Criteria priorities</th>
<th>r1</th>
<th>r2</th>
<th>r3</th>
<th>r4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance for the customer</td>
<td>0.615</td>
<td>0.263</td>
<td>0.203</td>
<td>0.267</td>
</tr>
<tr>
<td>Development time</td>
<td>0.292</td>
<td>0.311</td>
<td>0.311</td>
<td>0.198</td>
</tr>
<tr>
<td>Development complexity</td>
<td>0.093</td>
<td>0.266</td>
<td>0.263</td>
<td>0.276</td>
</tr>
<tr>
<td>The generalized priority</td>
<td>0.277</td>
<td>0.240</td>
<td>0.248</td>
<td>0.235</td>
</tr>
</tbody>
</table>

### Table 5 – Method AHP results

<table>
<thead>
<tr>
<th>№</th>
<th>The requirement (the initial text)</th>
<th>Keywords</th>
<th>Id.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>«The automated calculation system (ACC) should be expected to be used for processing the data about corporate customers, physical persons, legal bodies».</td>
<td>«The automated calculation system», &quot;customers&quot;, &quot;physical persons&quot;, &quot;legal bodies&quot;.</td>
<td>r1</td>
</tr>
<tr>
<td>2</td>
<td>«Creating complex tariff plans with the participation of services that belong to different personal accounts».</td>
<td>«Tariff plan», &quot;Personal account&quot;.</td>
<td>r2</td>
</tr>
<tr>
<td>3</td>
<td>&quot;The capability of working with groups of clients supporting functions &quot;friends&quot; and &quot;family&quot; and the cost-based user groups that are grouped based on arbitrary attributes, adding/ removing subscribers in a group as the operator and authorized person&quot;.</td>
<td>&quot;Customer&quot;, &quot;add/remove&quot;, &quot;subscribers&quot;.</td>
<td>r3</td>
</tr>
<tr>
<td>4</td>
<td>«Formation of the new, modification and removal of existing tariff plans by ACC operator».</td>
<td>&quot;Tariff plan&quot;, &quot;modification&quot;, &quot;removal&quot;.</td>
<td>r5</td>
</tr>
</tbody>
</table>

### Bibliography:

