Measuring Service Cohesion Using Latent Semantic Indexing

Automated Software Engineering Research Group
Electrical and Computer Engineering Faculty, Shahid Beheshti University GC,
Tehran, Iran
E-mail :{Ali.Kazemi,A.Rostampour}@mail.sbu.ac.ir
{F_Shams, P_Jamshidi}@sbu.ac.ir,Ali_Nasirzadeh@tabrizu.ac.ir

Abstract—As low coupling, high cohesion is a service-oriented
design and development principle that should be kept in mind
during all stages. High cohesion increases the clarity and ease
of comprehension of the design that simplifies maintenance
and achieves service granularity at a fairly reasonable level.
However, unlike coupling that only measures the degree of
structural and behavioral dependency to the other services,
cohesion metrics need to evaluate the degree of semantic
relationships between operations within a service in order
to measure functional relatedness. Latent semantic indexing
(LSI) is one of the techniques in the field of information
retrieval which is widely used to measure the degree of
semantic relatedness between a document and a given query
and also used to measure the cohesion of a text. In this paper,
we propose an approach to automatically measure the strength
of conceptual cohesion of a service based on LSI technique.
Finally, it has been evaluated theoretically based on a set of
cohesion principles.

Keywords—Service cohesion; Latent semantic indexing;
Software metric.

I. INTRODUCTION

Service-oriented architecture (SOA) is a promising
solution to build enterprise application programs which
supports processes and functions as a set of well-defined
services [1], [2]. Simply, a service is defined as a set of
related operations. Thus, it is a logic encapsulated by
individuals which supposed to be reusable [3]. Considering
design standards, which causes services to be potentially
reusable, the chance of a service to be able to accommodate
future requirements with the least development effort
increases [4]. Therefore, reusability is an important quality
attribute which must be measured to satisfy an important
need in SOA, the need for independent services to deliver a
reusable functionality [3], [4]. One of the design attributes
which has a great impact on reusability of a particular
service, is cohesion, so that higher cohesion significantly
increases service reusability [5], [6]. Also in [7], there is an
elaborate discussion about the impact of cohesion on
maintainability. The higher the cohesion of a service, the
easier the test and analysis and higher cohesion will improve
system stability and changeability [8] and, consequently
maintainability of a system will be improved.

Because of inherently conceptual nature, cohesion is one
of the most complicated and difficult structural attribute of a
class, component, or a service from quantifying point of
view [8]. This quality attribute can be measured based on
conceptual relatedness degree of operations which are
exposed in service interface. According to a cohesion
category proposed in [6], [7], conceptual cohesion is
considered as the strongest type of cohesion. However, this
type of service cohesion cannot be easily measured using
the previous traditional metrics due to additional level of
abstraction and highlighted characteristics of service
interfaces in comparison with procedural and object-oriented
paradigms [6].

Using the concept of latent semantic indexing (LSI), we
will evaluate the conceptual relatedness degree of operations
existing in a service. For the first time, LSI was used in
information retrieval techniques [9], [10]. One of the
applications of this method is measuring cohesion of a text
[11]. LSI provides completely automatic approach that
compares information units in order to measure conceptual
relatedness. Measuring conceptual relatedness degree of
units relies on a powerful mathematical method called
Singular Value Decomposition (SVD) [10], [11]. Therefore,
the objective of this research work is to propose a LSI-based
approach for measuring the degree of conceptual cohesion in
a service.

In order to adopt the LSI technique to measure service
cohesion, we get the required information of interactions
between business processes and business entities that are
mostly used during service identification [6]. The SVD
method is applied on a well-defined structure comprising this
information. The output of this algorithm is used to
quantitatively measure conceptual cohesion degree of the
identified services. Utilization of semantics existing in
technology-based enterprise processes is completely proportional to this
inherently conceptual nature, and therefore we will have a
more precise measurement of conceptual cohesion of
services.

The rest of this paper is organized as follows. Section II
introduces the most related works. In Section III, basic
concepts of the utilized terminologies are defined. The LSI
concepts and the way of adopting them are introduced in
Section IV. The proposed metric and complementary
example and issues are discussed in Sections V, VI and VII,
respectively. The theoretical principles of the metric are
evaluated in Section VIII. Finally, the conclusion, which leads to further research, is explained in Section IX.

II. RELATED WORK

In this section, we briefly present some of the previous works on measuring cohesion in service-oriented, object-oriented and procedural paradigms. The concept of cohesion in OO and procedural paradigms has been widely discussed and examined. For example, in [12], six semantic categories of procedural cohesion namely Coincidental, Logical, Temporal, Communicational, Sequential, and Functional have been proposed. The concept of cohesion later was extended by Eder et al. [13] to cover conceptual and technical features introduced in OO paradigm. Eder et al. [13] proposed five cohesion categories (from the weakest to strongest): Separable, Multifaced, Non-delegated, Concealed and Model. Moreover, in [6], eight semantic categories of service-oriented cohesion are proposed. These categories are: Coincidental, Logical, Temporal, Communicational, External, Implementation, Sequential, and Conceptual. In [6], four categories namely Communicational, External, Implementation, and Sequential are represented as quantifiable cohesion categories. On the other hand, four categories namely Coincidental, Logical, Temporal, and Conceptual are identified in this paper as purely semantic cohesion categories. They believe that second four categories are semantic based whereas first ones are measurable without considering semantic issues. The proposed quantifiable cohesion categories have indirect impact on conceptual cohesion. A brief representation of the cohesion metrics has been shown in Table I.

<table>
<thead>
<tr>
<th>Table I. SUMMARY OF COHESION METRICS IN THE LITERATURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>LCOM</td>
</tr>
<tr>
<td>LCOM3</td>
</tr>
<tr>
<td>RLCOM</td>
</tr>
<tr>
<td>TCC</td>
</tr>
<tr>
<td>WTCoh</td>
</tr>
<tr>
<td>SIDC</td>
</tr>
<tr>
<td>DM FAUM</td>
</tr>
<tr>
<td>SIDC</td>
</tr>
<tr>
<td>SIUC</td>
</tr>
<tr>
<td>SIC</td>
</tr>
<tr>
<td>TICS</td>
</tr>
</tbody>
</table>

It is worth to mention that in [20] and in [8], in addition to the number of shared parameters, other shared attributes such as number of service consumers, operations sequence and some more shared attributes are considered. To the best of our knowledge, there is no metric which measures the degree of relationship between operations of a service from conceptual point of view. Most proposed cohesion metrics in previous studies expect the services to have common inputs and outputs and does not consider the inter-relation of their parameters. To measure the conceptual cohesion, we require additional semantics. Therefore, we should look for methods that can measure the strength of conceptual relationship between two operations of a service by means of assets which are available in design level (processes from which services are identified) and then propose a metric for measuring conceptual cohesion of a service.

III. BASIC CONCEPTS

In this section, we present definitions of several key notions that will be utilized in this paper.

**Definition 1 (Business Entity):** A business entity (BE) is a dominant information entity with an associated data model and an associated behavior model in the context of a process scope [23].

**Definition 2 (Elementary Business Process):** An elementary business process (EBP) can be defined as EBP = {n, (BE, sr)}, where n is the name of elementary business process, BEj is the jth business entity which semantically relate to corresponding EBP. sr = {"C", "R", "U", "D"} is the type of semantic relationship between EBP and BEj[22].

**Definition 3 (CRUD matrix):** A CRUD matrix can be defined as M= {(EBPj, BEi) i=1...#row, j=1...#column}, where EBPi is the i-th EBP and BEj is the j-th BE. #row is the number of EBP and #column is the number of BE in the model [22].

**Definition 4 (Conceptual Cohesion):** There is a meaningful semantic relationship between all operations of a service in terms of some identifiable domain-level concept. [6].

IV. APPLICABILITY OF LSI IN COHESION MEASUREMENT

LSI is a vector model-based technique which is applied in many information retrieval applications. In the vector model, each document is simply represented by a \( A_{n \times m} \) term-document matrix, where \( n \) is the number of terms and \( m \) is the number of documents in the collection. Each cell \( a_{ij} \) is the frequency of term \( t_i \) in the document \( d_j \). LSI technique includes the following main steps:

1. A matrix is formed; each row of this matrix is corresponded to a term which occurs in the document. Each element (m,n) in the matrix is corresponded to number of times that term m occurs in document.
2. Local and global weighting of terms is applied to each element of the term-document.
3. SVD is used by LSI and decomposes the matrix into three other matrices: \( T \), a term in the dimension; \( S \), a
diagonal matrix of singular values, and D a document matrix in the dimension. The number of dimensions is considered as \( t = \min(m, n) \) where \( m \) and \( n \) are the number of the terms and the number of documents in the main term-document matrix respectively. The matrix can be provided by \( A = T S D^T \) where \( D^T \) is transposed of matrix \( D \).

4. In the LSI system, the T, S and D matrices are truncated to \( k \) dimensions. Dimensional reduction reduces “noise” in the term–term matrix resulting in a richer term relationship structure that reveals latent semantics and is a crucial step in this research work.

Now we explain each one of the above steps in more details.

In the first step, the term-document matrix A is formed.

In the second step, a weight is assigned to each term in the document. There are different weighting models which are explained in [20][9]. The simplest weighting model can be obtained simply by counting number of frequency of a term in the document. In order to put the weights in the interval \([0,1]\), the weight of each term is divided in document by \( t_f_{\text{max}} \), where \( t_f_{\text{max}} \) is the maximum of term in the document.

In the third step, term-document matrix, A is taken and then is decomposed into three matrices T, S, and D using SVD. Matrices T, S and D keep the information related to terms, singular values and documents respectively.

In the fourth step, T, S, and D matrices are decreased to \( K \) domains. After dimensional reduction, the term-term matrix can be approximated using the formula: \( TTS = T K S K (T K S K)^T \).

In this work, we suppose that the value which exists in location \((i,j)\) of \( TTS \) matrix show the similarity between terms \( i \) and \( j \) in the collection. The value of \( K \) is optional, in this paper according to [20], \( K=2 \). Our main goal in this paper is to present cohesion metric which is able to measure the strength of conceptual similarity between operations of a service.

As we mentioned in Section IV, to measure a service cohesion using the proposed metric, first the BE-EBP matrix should be formed. This matrix can be formed based on those enterprise processes which services are intended to be obtained from their decomposition as defined in definition 3.

The measuring procedure has the following form. Firstly a matrix \( A_{nxm} \) is formed where \( m \) is the number of enterprise business entities and \( n \) is the number of enterprise processes. Then the number of times that each business entity \( i \) accessed by business process \( j \), is considered as element \((i,j)\) of matrix \( A \). In order to obtain conceptual relatedness between business entities, we apply SVD on matrix \( A \). Its outputs are three matrices which are shown as \( A = T S D^T \).

As we discussed earlier, \( TTS = T_2 S_2 (T_2 S_2)^T \) matrix shows the conceptual relatedness between business entities which are used to obtain the strength of service cohesion. For this purpose, we use a graph based approach.

Suppose that service \( S \) has a set of operations \( O = \{ O_1, O_2, ..., O_m \} \). Each operation \( O_i \) of the service \( S \) accesses a set of business entities which is shown as \( BE_j = \{ BE_{j,1}, BE_{j,2}, ..., BE_{j,n} \} \). For each pair of operations \( O_i \) and \( O_j \) in the service \( S \) we form a complete graph \( G = (V,E) \) so that \( V = BE_i \cup BE_j \).

Now, in set \( E \), we assign a value for each edge that represents the degree of relationship between business entities, which is considered as nodes in graph \( G \). The degree of relationship between two business entities can be measured from \( TTS \) matrix. The degree of conceptual
relatedness between two operations \(i\) and \(j\) is calculated through formula:

\[
OCV(i, j) = \begin{cases} 
\frac{\sum_{p,q} \sum_{p,q} TTS_{p,q}}{|V|} \quad & |V| > 1 \\
1 \quad & |V| = 1
\end{cases}
\]

where:
- \(p\) and \(q\) are two business entities in \(V\).
- \(TTS_{p,q}\) is the degree of relationship between two business entities, \(BE_p\) and \(BE_q\).
- \(|V|\) is the cardinality of set \(V\).
- The denominator is the number of edges in the complete graph \(G\).

The strength of cohesion is defined as the degree of relationship between service's operations.

\[
SCV(S) = \begin{cases} 
\frac{\sum_{v \in V} \sum_{v \in V} OCV(i, j)}{m^2 (m-1)/2} \quad & |m| > 1 \\
1 \quad & |m| = 1
\end{cases}
\]

where:
- \(m\) is the number of operations in service \(S\).

VI. EXAMPLE

In this section, we show how the proposed metric works using an example. To do that, we must have the enterprise processes and services which are identified using those processes. Using a real-world business process the effectiveness of the proposed metric is studied and evaluated. The sales department is studied in this scenario [22].

Using CRUD matrix is one of the ways to identify a service [22]. Table III illustrates the CRUD matrix associated to our scenario. Identified services are shown in the form of clusters with different colors (Table III).

V. Example

In this section, we show how the proposed metric works using an example. To do that, we must have the enterprise processes and services which are identified using those processes. Using a real-world business process the effectiveness of the proposed metric is studied and evaluated. The sales department is studied in this scenario [22]. Using CRUD matrix is one of the ways to identify a service [22]. Table III illustrates the CRUD matrix associated to our scenario. Identified services are shown in the form of clusters with different colors (Table III).

The BE-EBP matrix is shown in Table IV. Since EBPs existing in the CRUD matrix access each business entities just 0 or 1 time, elements of this matrix are just 0 and 1. For example, Add Customer accesses only customer and credit BEs, therefore there are just two ones in Add Customer column.

After the matrix of BE-EBP is obtained, we apply SVD algorithm on it. To do that, MATLAB version 7.6.0.324 has been used. To obtain business process entity matrix we use this equation:

\[
TTS = T_2S_2(T_2S_2)^T
\]

The resulted matrix has been shown in Table V. Also this matrix has been normalized and its negative values have been substitute with 0s.

Next we show how to calculate the cohesion of a service using the proposed metric. Table III show a CRUD matrix with four identified services. First we show how to calculate the metrics for the first service which is shown by blue color.

We have:

\[
O = \{O_1, O_2, O_3\}
\]

\[
BE_1 = \{Customer, Credit\}
\]

\[
BE_2 = \{Customer, Credit, Accountreceivable\note\}
\]

\[
BE_3 = \{Customer, Credit\}
\]

In order to obtain conceptual relatedness between operations of a service we use a graph. For operations \(O_1\) and \(O_2\), graph \(G= (V,E)\) has the form of Figure 1. In this graph the set \(V\) has the following form.
$V = BE_1 \cup BE_2 =\{Customer, Credit, Accountreceivable, note\}$

Figure 1. Business Entities Graph for Service 1

$OCV(1,2) = \frac{0.9967 + 0.1954 + 0.2712}{3} = 0.4877$

The results of the alternatives for the specified service $S_1$ is shown in Table VI.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$O_iO_j$</th>
<th>$O_iO_k$</th>
<th>$O_iO_l$</th>
<th>$O_jO_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$OCV$</td>
<td>0.4877</td>
<td>0.9967</td>
<td>0.4877</td>
<td></td>
</tr>
</tbody>
</table>

Finally, the strength of conceptual cohesion of service is obtained.

$SCV(S_1) = \frac{0.4877 + 0.9967 + 0.4877}{3} = 0.6573$

Table VII shows the conceptual cohesion of four identified services on CRUD matrix of Table III.

<table>
<thead>
<tr>
<th>Service</th>
<th>The value of cohesion (SCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.6573</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.6256</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.2658</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.0524</td>
</tr>
</tbody>
</table>

VII. DISCUSSION

The results clearly demonstrate that our proposed metric for cohesion appropriately measure conceptual cohesion of a service. Now, we analyze the values which provided by the proposed metric and the previous metrics such as SIDC [6] and TCC [17] and CCM [22]. As explained earlier, the operations of a service must be related in terms of some domain-level concepts. In other words, they must be focused on single business functionality. The analysis shows that semantics in business process are utilized properly in the proposed metric, so it evaluates service cohesion completely in conceptual point of view.

We consider two identified services in CRUD matrix (Table III), highlighted with red and blue colors, as the material for analysis. Each of these services has three operations; their operations and the resulted cohesion value, obtained by three mentioned metrics have been shown in Table VIII. Values shown in Table VIII state the relationship between two operations of each service. Consider group B1 of the first service and group R3 of the second service. For B1 and R3 groups, the SIDC and TCC give the same result value whereas these groups have different cohesion in conceptual point of view.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Group</th>
<th>EBP Index</th>
<th>Proposed Metric</th>
<th>SIDC</th>
<th>TCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>B1</td>
<td>1,2</td>
<td>0.48</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>Red</td>
<td>R3</td>
<td>10,11</td>
<td>0.21</td>
<td>0.66</td>
<td>0.66</td>
</tr>
</tbody>
</table>

In B1, Credit and Customer BEs have been accessed together four times (by 1, 2, 3, 9 EBPs of CRUD matrix), these two entities are very related conceptually, because according to [22] two BEs are related if there is at least one shared activity in their behavioral model. In behavioral model of Credit and Customer, there are four shared activities that are processed simultaneously. Therefore, any action on one of them requires an action on the other. In other words, in this organization, whenever an operation performed on Customer, we can expect that an operation must be done on Credit entity. Thus generally, we can associate performing an action on one of them with performing an action on the other as an atomic activity (Create, Update, Read and Deletion of Credit entity and vice versa). On the other hand, existing high cohesion between service operations can be considered as a reusability predictor. This capability is provided by the proposed metric clearly. It is obvious that whenever an operation on a BE is performed with another operation on another BE frequently in enterprise processes, means that performing these two operations together has higher potential reusability. Consequently, it is better to place these two operations, which are considered as an atomic activity, in the same service. Existing of Account Receivable Note in this service (Blue Cluster) results in corruption of this service. Because this BE shares one activity in its behavioral model with behavioral model of other two entities (second EBP of CRUD matrix). Therefore, the cohesion value of 0.48 has been obtained for group B1. In R3 group, two BEs, Issue and Shipping, have been accessed together just two times (by 10,11 EBPs of CRUD matrix), so we can say that these entities are less related in conceptual point of view in comparison with Credit and Customer entities. Moreover, Draft has just one shared activity with two other entities in its behavioral model (EBP11 of CRUD matrix). By conducting similar analysis, there is a lower cohesion between existing operations in R3 in comparison with B2 which our metric shows this point by obtaining cohesion value of 0.21.

VIII. EVALUATION OF THE METRIC

The proposed cohesion metric is analytically evaluated by using property-based software engineering measurement framework [24]. The metric satisfies all of the cohesion properties and therefore it can be a valid measure of cohesion from the measurement theory point of view.

Property 1: Non-negativity and Normalization are satisfied because SCV metric never becomes negative under any conditions, and its value will be in [0,1]. Normalization always let the direct and meaningful comparison between strength of services' cohesion.

Property 2: Null Value is satisfied because SCV metric gets the value if mutual relationship between all of business entities which are used by operations of a service is 0.
Property 3: Monotonicity is satisfied because by adding a related business entity to a pair EBP its overall cohesion is not decreased. In the other words, whenever we add a related business entity to a set of BEs which are accessed by a pair service operations, the cohesion between those two operations will not be decreased.

Property 4: Cohesive Modules is satisfied because by joining two unrelated service interface, the resulted cohesion will not be greater than the cohesion of original interfaces. In the other words, the strength of cohesion between operations of two unrelated services will not be greater than the strength of each service, because they access unrelated BEs.

IX. Conclusion and Future Work

In this paper, using LSI technique, the strength of conceptual cohesion of a service was measured. In this technique, the business entity-business process matrix is formed using existing semantics in business processes and then by applying SVD algorithm on this matrix, the business entity-business entity matrix was resulted so that this matrix represented conceptual relationship between business entities. By adopting business entity-business entity matrix, we can measure the strength of conceptual cohesion of candidate services in the service identification phase. This quality attribute has a great impact on service reusability and maintainability but inherently conceptual nature caused it to be very difficult from quantifying point of view. Therefore measuring this important quality attribute from the conceptually point of view is very valuable. Writers' lack of access to all enterprise processes caused that the effectiveness of the cohesion measuring approach to be shown using a CRUD matrix. Although the obtained results approves the effectiveness of proposed metric well, but having all processes of a real enterprise and then using this metric in the service identification phase completely approves usefulness of this metric. Therefore, using more case studies can be considered as future work of this paper.

Acknowledgment

The project has been partially supported by Iran Education and Research Institute for Information and Communication Technology (formerly Iran Telecommunication Research Center (ITRC)) under contract number 4290/500 and also Shahid Beheshti University under the supervision Automated Software Engineering Research (ASER) Group, Faculty of Electrical and Computer Engineering.

References