From Network Traces to System Responses: Opaquely Emulating Software Services

Miao Du*, Steve Versteeg†, Jean-Guy Schneider*, John Grundy*, Jun Han*
*School of Software and Electrical Engineering
Swinburne University of Technology, Hawthorn, Victoria 3122, Australia
Email: {miaodu,jschneider,jgrundy,jhan}@swin.edu.au
†CA Technologies, Melbourne, Victoria 3004, Australia
Email: steve.versteeg@ca.com

Abstract—Enterprise software systems make complex interactions with other services in their environment. Developing and testing for production-like conditions is therefore a challenging task. Prior approaches include emulations of the dependency services using either explicit modelling or record-and-replay approaches. Models require deep knowledge of the target services while record-and-replay is limited in accuracy. We present a new technique that improves the accuracy of record-and-replay approaches, without requiring prior knowledge of the services. The approach uses multiple sequence alignment to derive message prototypes from recorded system interactions and a scheme to match incoming request messages against message prototypes to generate response messages. We introduce a modified Needleman-Wunsch algorithm for distance calculation during message matching, wildcards in message prototypes for high variability sections, and entropy-based weightings in distance calculations for increased accuracy. Combined, our new approach has shown greater than 99% accuracy for four evaluated enterprise system messaging protocols.

I. INTRODUCTION

Software systems are becoming ever increasing interconnected. Service emulation supports developing and testing a software system, independent of the other systems on which it depends [1]. In a typical deployment scenario, an enterprise system might interact with many other systems, such as a mainframe, directory servers, databases and other types of software services. In order to check whether the enterprise system will function correctly in terms of its interactions with these services, it is necessary to test it in as realistic an environment as possible, prior to the actual deployment. Getting access to the actual production environment for testing is not possible due to the risk of disruption. Large organisations often have a test environment, which is a close replication of their production environment, but this is very expensive. Furthermore the test environment is in high demand, so software developers will have only limited access to it. Enabling developers to have continuous access to production-like conditions to test their application is an important part of Development-Operations (known as ‘DevOps’) [2][3].

One popular approach that developers use to test their application’s dependence on other systems is to install the other systems on virtual machines (such as VMware) [4]. However virtual machines are time consuming to configure and maintain. Furthermore the configuration of the systems running on the virtual machine is likely to be different to the production environment. An alternative that is gaining traction is service emulation, where models of services are emulated - sometimes into the many thousands of service instances - to provide more realistic scale and less complicated configuration [5]. However, existing approaches to service emulation rely on system experts explicitly modelling the target services and this requires very detailed knowledge of message protocols and structure. This is often infeasible if the required knowledge is unavailable for the wide range of services, especially legacy services, in a real deployment environment [6], and is besides very time consuming and error-prone.

Our aim is to develop an automated approach to service emulation which uses no explicit knowledge of the services, their message protocols and structures and yet can simulate – to a high degree of accuracy and scale – a realistic enterprise system deployment environment. To achieve this we need an automated, accurate, efficient and robust method for service emulation that derives service responses from collected actual network traces between an enterprise system and a service – the focus of this paper.

Our previous work [7] relied on searching the collected message traces to automatically generate service responses. Unfortunately this was limited to text based protocols and became inefficient for large message traces. The method we employed was sensitive to mismatching by payload versus operation type (i.e. was not robust.) Our subsequent work [8] used clustering to improve efficiency but decreased accuracy. The aim of the research described in this paper is to achieve both higher accuracy and efficiency through the use of descriptive and robust message prototypes which can be applied to both binary and text protocols. Our approach uses:

- a multiple sequence alignment algorithm to derive message prototypes,
- wildcards in message prototypes for the sections with high variability,
- a modified Needleman-Wunsch algorithm for message distance calculations, and
- entropy-based weightings in distance calculations for increased accuracy.

We present a set of experiments with four enterprise system
messaging protocols, including binary and textual LDAP, a binary Mainframe protocol, a textual JSON (Twitter) and SOAP services. These experiments show overall a greater than 99% accuracy in the generated response messages for the four protocols tested. Additionally they show efficient performance in generating the emulated service responses, enabling scaling within an emulated deployment environment.

II. MOTIVATION AND RELATED WORK

Consider an example enterprise software system which is to be integrated with other systems in its production environment as part of its operation. The other systems (called services) include a legacy mainframe program, a directory server and a web service. The behaviour of the example enterprise software system depends on the responses it receives from these dependency services. The directory server uses a proprietary protocol which is poorly documented. It has operation types add, search, modify and delete, each with various payloads and response messages. The operation type is encoded in a single character. A small sample set of request and response messages (a transaction library) for the directory service is shown in Table I. Such message traces are captured by either network monitoring tools, such as Wireshark [9], or proxies. In the following sections we use this sample transaction library to illustrate how our method works. This transaction library is from a fictional protocol that has some similarities to the widely used LDAP protocol [10], but is simplified to make our running example easier to follow.

The enterprise system under test needs to interact with this service, to search for user identities, register new identities, de-register identities and so on. Other systems in the environment share the same service. For complex deployment environments we want to extensively test the enterprise system as to its protocol conformance, robustness and scaling.

The most common approach to providing a testing environment for enterprise systems using such a messaging protocol is to use virtual machines [11]. Implementations of the services are deployed on virtual machines and communicated with by the system under test. Major challenges with this approach include configuration complexity [12] and the need to maintain instances of each and every service type in multiple configurations [13]. Recently, cloud-based testing environments have emerged to mitigate some of these issues [14].

Emulated testing environments for enterprise systems, relying on service models, are another approach. When sent messages by the enterprise system under test, these respond with approximations of “real” service response messages [15]. Kaluta [16] is proposed to provision emulated testing environments. Challenges with these approaches include developing the models, lack of precision in the models, especially for complex protocols, and ensuring robustness of the models under diverse loading conditions [17]. To assist developing reusable service models, approaches either reverse engineer message structures [18][19][20], or discover [21][22] processes for building behavioral models [23]. While these allow engineers to develop more precise models, none of them can automate interaction between enterprise system under test and the emulated dependency services.

Recording and replaying message traces is an alternative approach. This involves recording request messages sent by the enterprise system under test to real services and response messages from these services, and then using these message traces to ‘mimic’ the real service response messages in the emulation environment [24]. Some approaches combine record-and-replay with reverse-engineered service models. CA Service Virtualization [25] is a commercial software tool, which can emulate the behaviour of services. The tool assumes knowledge of protocol message structures to model services and mimic interactions automatically.

III. APPROACH

Our goal is to produce an emulation environment for enterprise system testing that uses message trace recordings collected a priori to produce a response on behalf of a service when invoked by a system-under-test at runtime.

The general approach is to cluster the trace recordings into groups of similar messages and then formulate a single representation for the request messages of each cluster. This accelerates runtime performance by enabling incoming requests from the system under test to be compared only to the cluster representations, rather than the entire transaction library. Previous work [8] selected the cluster centroid request as the representative. However accuracy decreased as the information from the other requests in the cluster was discarded. Our current aim is to improve accuracy, but preserve efficiency, by generating prototypes which capture the common features of the range of requests in each cluster.

1 The centroid is the transaction with the minimised total distance from the other transactions in the cluster.
The approach for generating and using our prototypes (depicted in figure [1]) has the following steps:

1) We collect message traces of communications between a client and the real target service and store them in a transaction library with a transaction being a request-response pair.

2) We then cluster the transaction library, with the goal of grouping transactions by operation type. We do not consider the state of service under which the requests are issued. (This may give lower accuracy but is still useful in many cases as discussed in Section [7].)

3) We then derive a request consensus prototype (a message pattern) for each operation type cluster by:
   - a) Aligning all request messages of each cluster to reveal their common features for each operation type by inserting gap characters;
   - b) Extracting a consensus character sequence from the aligned request messages (i.e., the request message prototype) by selecting or deleting a character at each character position;
   - c) Calculating positional weightings to prioritise different sections of each request consensus prototype.

4) At runtime, for an incoming request message, we use a matching distance calculation technique to select the nearest matching request consensus prototype.

5) We then perform dynamic substitutions on a specifically chosen response message from the identified operation type cluster to generate a final response message to be sent back to the enterprise system under test.

Steps 1 to 3 are performed offline to prepare request consensus prototypes. Steps 4 and 5 are performed at runtime to generate live responses to the system under test messages.

### A. Needleman-Wunsch

The Needleman-Wunsch algorithm [26] is used at different steps during the offline and runtime processing. It is a dynamic programming algorithm which finds the globally optimal alignment for two sequences of symbols in $O(MN)$ time, where $M$ and $N$ are the lengths of the sequences. Needleman-Wunsch uses a progressive scoring function $S$, which gives an incremental score for each pair of symbols in the alignment. A different score will be given depending on whether the symbols are identical, different or a gap has been inserted.

### B. Collecting and Clustering Interaction Messages

The first step is to record real message exchanges between a client (such as a previous version of the system under test) and the service we aim to emulate. A transaction is a request-response pair of the communications, where the service may respond with zero or more response messages to a request message from the client. We record requests and responses at the network level (using a tool such as Wireshark), recording the bytes in TCP packet payloads. This makes no assumptions about the message format of the service application.

Having recorded our transaction library, the next step is to group the transactions by operation type, but again without assuming any knowledge of the message format. To achieve this we use a distance function-based clustering technique. In our previous work [8] we considered multiple cluster distance functions and found that the response similarity (as measured by edit distance, calculated using the Needleman-Wunsch algorithm [26]) was the most effective method for grouping transactions of the same operation type. We therefore use the same technique in this work, grouping transactions by the similarity of their response messages. The clustering algorithm used was VAT [27], one among many alternatives, as we found it to be effective in our previous work [8].

Applying our clustering technique to the example transaction library yields two clusters, as shown in tables [1] and [II] corresponding to search operations and add operations, respectively.

### C. Aligning Request Messages

To formulate a representative cluster prototype, we first align the request messages in a cluster to determine the common features of these request messages. Then we extract the common features while accommodating variations. In aligning the request messages of a cluster, we adopt the multiple sequence alignment (MSA) technique originated from Biology. MSA was first used to align three or more biological sequences to reveal their structural commonalities [28]. Using MSA for revealing commonalities of interaction messages offers a number of advantages over other techniques, such as $n$-grams [31]. Firstly, it is able to effectively handle messaging protocols with single-byte operation fields (e.g. in the binary LDAP protocol) as well as multi-byte operation fields. Secondly, it does not

---

<table>
<thead>
<tr>
<th>Index</th>
<th>Request</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{id:1,op:S,sn:Du}</td>
<td>{id:1,op:SearchRsp,result:Ok,gn:Miao,sn:Du,mobile:5362634}</td>
</tr>
<tr>
<td>24</td>
<td>{id:24,op:A,sn:Schneider,mobile:123456}</td>
<td>{id:24,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>275</td>
<td>{id:275,op:S,sn:Han}</td>
<td>{id:275,op:SearchRsp,result:Ok,gn:Jun,sn:Han,mobile:33533333}</td>
</tr>
<tr>
<td>2273</td>
<td>{id:2273,op:S,sn:Schneider}</td>
<td>{id:2273,op:SearchRsp,result:Ok,sn:Schneider,mobile:123456}</td>
</tr>
<tr>
<td>2487</td>
<td>{id:2487,op:A,sn:Will}</td>
<td>{id:2487,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>3106</td>
<td>{id:3106,op:A,sn:Hine,gn:Cam,Postcode:33589}</td>
<td>{id:3106,op:AddRsp,result:Ok}</td>
</tr>
</tbody>
</table>

TABLE I: Directory Service Interaction Library Example

---

2For this reason this technique has also been widely used to reverse-engineer protocol message structures [29], [30].
require a predetermined matching sequence width (e.g., \( n \)) in the \( n \)-gram.

In particular, we adopt ClustalW [32], a widely used heuristic technique for MSA. It is memory efficient and is shown to produce high accuracy alignments in polynomial computation time for empirical datasets (in contrast to the original NP-complete MSA technique [33]). ClustalW is a progressive MSA algorithm, where pairwise sequence alignment results are iteratively integrated into the multiple sequence alignment result. An overview of the ClustalW algorithm is as follows:

1) All \( N(N-1)/2 \) pairs of sequences are aligned to calculate their similarity ratio using the Needleman-Wunsch algorithm, where \( N \) is the number of sequences.
2) A \( N \times N \) distance matrix is built for capturing distance calculation.
3) A guide tree is constructed from the distance matrix by applying a neighbour-joining clustering algorithm. (A guide tree is a tree data structure which organises the similarities between sequences.)
4) The guide tree is used to guide a progressive alignment of sequences from the leaves to the root of the tree.

Figure 2 shows the multiple sequence alignment results of applying the ClustalW algorithm to the example clusters from tables II and III. The MSA results are known as profiles. Gaps which were inserted during the alignment process are denoted with the ‘\( \ast \)’ symbol. Note that the common sequences for the requests in each cluster have now been aligned.

![Figure 2: MSA results of the requests in tables II and III](image)

### D. Formulating the Request Consensus Prototype

Having derived the MSA profile for the request messages of each cluster, the next step is to extract the common features from the MSA profile into a single character sequence, which we call the request consensus prototype, to facilitate efficient runtime comparison with an incoming request message from the system under test.

From all the aligned request messages in a cluster, we derive a byte (or character) occurrence count table. Figure 3 graphically depicts byte frequencies at each position for the example alignment in Figure 2. Each column represents a position in the alignment result. The frequencies of the different bytes in the MSA profile are displayed as a stacked bar graph. Based on the byte occurrence table, we formulate the request consensus prototype by extending the concept of a consensus sequence [34] commonly used in summarising a MSA profile. A consensus sequence can be viewed as a sequence of consensus symbols, where the consensus symbol \( c_i \) is the most commonly occurring byte at the position \( i \). In our extension, a request consensus prototype, \( p_i \), is calculated by iterating each byte position of the MSA profile, to calculate a prototype symbol, \( p_i \), at each position, according to equation 1.

\[
p_i = \begin{cases} 
  c_i & \text{if } q(c_i) \geq f \land c_i \neq \ast \\
  \bot & \text{if } q(c_i) \geq \frac{1}{2} \land c_i = \ast \\
  ? & \text{otherwise}
\end{cases}
\]  

(1)

Where \( q(c_i) \) denotes the relative frequency at position \( i \) of the consensus symbol \( c_i \), \( f \) is the relative frequency threshold, ‘\( \ast \)’ denotes a gap, ‘?’ is the ‘wildcard’ symbol and ‘\( \bot \)’ represents a truncation. After calculating the prototype symbol for each position, any truncation symbols are then deleted from the request consensus prototype.

Introducing wildcards and truncations into the message prototypes allows us to distinguish between gaps and where there is no consensus. If the relative frequency \( q(c_i) \) is at or above the threshold \( f \), then we insert the consensus symbol

<table>
<thead>
<tr>
<th>Index</th>
<th>Request</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{id:1,op:S,sn:Du}</td>
<td>{id:1,op:SearchRsp,result:Ok,gn:Miao,sn:Du,mobile:5362634}</td>
</tr>
<tr>
<td>275</td>
<td>{id:275,op:S,sn:Han}</td>
<td>{id:275,op:SearchRsp,result:Ok,gn:Jin,sn:Han,mobile:33333333}</td>
</tr>
<tr>
<td>2273</td>
<td>{id:2273,op:S,sn:Schneider}</td>
<td>{id:2273,op:SearchRsp,result:Ok,sn:Schneider,mobile:123456}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>Request</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>{id:24,op:A,sn:Schneider,mobile:123456}</td>
<td>{id:24,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>2487</td>
<td>{id:2487,op:A,sn:Will}</td>
<td>{id:2487,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>3106</td>
<td>{id:3106,op:A,sn:Hin,gn:Cameron,postalCode:3*3589}</td>
<td>{id:3106,op:AddRsp,result:Ok}</td>
</tr>
</tbody>
</table>

### TABLE II: Cluster 1 (search operations)

### TABLE III: Cluster 2 (add operations)
into our prototype (unless the consensus symbol is a gap.) If the relative frequency is below threshold, then we insert a wildcard. If the consensus symbol is a gap and it is in the majority, then we leave that position as empty (i.e. deleted). Wildcards allow us to encode where there are high variability sections of the message. In our experimentation, without truncations consensus sequences became artificially long as there tended to be many gaps. By truncating the gaps, the lengths of the prototypes become similar to the typical lengths of messages in the cluster. The consensus prototype for a cluster of request messages can differentiate stable positions from variant positions. Moreover, it can identify consensus symbols that can be utilised for matching.

Applying our request consensus prototype method, using a frequency threshold \( f = 0.8 \), to the example clusters in tables II and III yields the following results:

**Consensus prototype for the search request cluster:**

\{id:????,op:S,sn:???????}\n
**Consensus prototype for the add request cluster:**

\{id:????,op:A,sn:????????????????????????\}

(Note that the add prototype contains an ‘I’ from coincidentally aligning ‘I’ from ‘mobile’, ‘Will’ and ‘postalCode’.)

**E. Deriving Entropy-Based Positional Weightings**

The final step of the offline analysis is to calculate weightings for each byte position in the consensus prototype to support the runtime distance matching process for request messages. In our service emulator, generating a response message of the correct operation type is more critical than the contents of the message payload. Thus we give a higher weighting to the sections of the message that likely relate to the operation type. To do this we make use of the observation that structure information (such as operation type) is more stable than payload information. We use entropy as a measurement of variability, and use it as the basis to calculate a weighting for each byte position of the consensus prototype.

Using the MSA profile for a cluster (from section III-C), we calculate the entropy for each column using the Shannon Index \( I \), as given by equation 2

\[ E_i = - \sum_{j=1}^{R} q_{ij} \log q_{ij} \]  

Where \( E_i \) is the Shannon Index for the \( i \)th column, \( q_{ij} \) is the relative frequency of the \( j \)th character in the character set at column \( i \) and \( R \) is the total number of characters in the character set.

Since we wish to give a high weighting to stable parts of the message, for each column we invert the entropy by applying a scaling function of the form given in equation 3

\[ w_i = \frac{1}{(1 + bE_i)^c} \]  

Where \( w_i \) is the weighting for the \( i \)th column, \( E_i \) is the entropy of the \( i \)th column, and \( b \) and \( c \) are positive constants. The higher the values of \( b \) and \( c \), the more higher entropy columns are deweighted. In our experiments we found the best results were obtained with \( b = 1 \) and \( c = 10 \). This allows structural information to strongly dominate in the matching process. Payload similarity is only used as a ‘tie-breaker’.

Columns that correspond to gaps removed from the consensus prototype are also dropped from the weightings array. Table IV gives an example weightings array for the search request consensus prototype.

**F. Entropy Weighted Runtime Matching for Request Messages**

At runtime, the formulated request consensus prototypes are used to match incoming requests from the system under test. We extend the Needleman-Wunsch algorithm to calculate the matching distance between an incoming request and the request consensus prototype for each operation type (or cluster). We modify the Needleman-Wunsch scoring function, \( S \) (see equation 4), by giving a special score for alignments with wildcard characters and multiplying all scores by their corresponding positional entropy weights (from equation 3).

\[ S(p_i, r_j) = \begin{cases} 
  w_i M & \text{if } p_i = r_j \land p_i \neq \? \\
  w_i D & \text{if } p_i \neq r_j \land p_i \neq \? \\
  w_i X & \text{if } p_i = \?
\end{cases} \]  

Where \( p_i \) is the \( i \)th character in the consensus prototype, \( r_j \) is the \( j \)th character in the incoming request, \( w_i \) is the weighting for the \( i \)th column, \( M \) and \( D \) are constants denoting the Needleman-Wunsch identical score and difference penalty, respectively, and \( X \) is the wildcard matching constant. In our experiments we used values \( M = 1 \), \( D = -1 \) and \( X = 0 \).

Using the modified scoring equation 4 we apply Needleman-Wunsch to align the consensus prototype with an incoming request, calculating an absolute alignment score, \( s \).

The relative distance, denoted as \( d_{rel} \), is calculated from the absolute alignment score to normalise for consensus prototypes of different lengths, different entropy weights and differ-
ent numbers of wildcards. The relative distance is in the range 0 to 1, inclusive, where 0 signifies the best possible match with the consensus, and 1 represents the furthest possible distance. It is calculated according to equation 5

$$d_{rel}(p, r) = 1 - \frac{s(p, r) - s_{min}(p)}{s_{max}(p)}$$  

(5)

Where $s_{max}$ is the maximum possible alignment score for the given consensus prototype and $s_{min}$ is the minimum possible alignment score for the given consensus prototype. These are calculated in equations 6 and 7 respectively.

$$s_{max}(p) = \sum_{i=1}^{n} S(p_i, p_i)$$  

(6)

$$s_{min}(p) = \sum_{i=1}^{n} S(p_i, \emptyset)$$  

(7)

Where $\emptyset$ is a special symbol, different to all of the characters in the consensus prototype.

The consensus prototype that gives the least distance to the incoming message is selected as the matching prototype, therefore identifying the matching transaction cluster.

As an example, suppose we receive an incoming add request with the byte sequence \{id:37,op:A,sn:Durand\}. Aligning the request against the search consensus prototype yields:

- **request:** \{id:37,op:A,sn:Durand\}
- **prototype:** \{id:????,op:S,sn:????????????\}

Using equation 5, the weighted relative distance is calculated to be 0.0715. In comparison, the add consensus prototype produces the alignment below with a relative distance of 0.068.

- **request:** \{id:37*',op:A,sn:*******D*u*r*an*d*+'\}
- **prototype:** \{id:???,op:S,sn:??????????????????????\}

Consequently, the add prototype is the closest matching, causing the add cluster to be selected.

### G. Response Generation through Dynamic Substitution

The final step in our approach is to send a customised response for the incoming request by performing some dynamic substitutions on a response message from the matched cluster. Here we use the symmetric field technique described in [7], where character sequences which occur in both the request and response messages of the chosen transaction are substituted with the corresponding characters from the live request in the generated response. We take the response from the centroid transaction [3] of the selected cluster and apply the symmetric field substitution.

In our example, the centroid transaction from the add cluster is given below. There is one symmetric field (boxed).

```
request: \{id:24,op:A,sn:Schneider,mobile:123456\}
response: \{id:37,op:AddRsp,result:Ok\}
```

After performing the symmetric field substitution the final generated response is: \{id:37,op:AddRsp,result:Ok\}

### H. Implementation

We implemented our proposed system using Java, following the steps in the previous sections, using the architecture of Figure 1. Due to commercial agreements we are unable to release the source code of our implementation.

## IV. Evaluation

We wanted to assess three key characteristics of our approach: **accuracy**, **efficiency** and **robustness**, aiming to answer the following research questions:

1. **RQ1 (Accuracy):** Having request consensus prototypes formulated at the pre-processing stage, is our approach able to generate accurate, protocol-conformant responses? At runtime, can **positional weighting based matching** further improve the accuracy of our approach for generating such responses?
2. **RQ2 (Efficiency):** Is our technique efficient enough to generate timely responses, even for large transaction libraries?
3. **RQ3 (Robustness):** Given non-homogeneous clusters, which contain some messages of operation types different to the majority of messages, is our technique robust enough to generate accurate responses?

In order to answer these questions, we applied our technique on six message trace datasets of four case study protocols. We ran the technique on these datasets and assessed its accuracy, efficiency and robustness for each. We then compared these results to our previous techniques.

### A. Case Study Protocols and Traces

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Binary/Text</th>
<th>Fields</th>
<th>#Ops.</th>
<th>#Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS [37]</td>
<td>binary</td>
<td>fixed length</td>
<td>5</td>
<td>800</td>
</tr>
<tr>
<td>LDAP</td>
<td>binary</td>
<td>length-encoded</td>
<td>10</td>
<td>2177</td>
</tr>
<tr>
<td>LDAP text (1)</td>
<td>text</td>
<td>delimited</td>
<td>10</td>
<td>2177</td>
</tr>
<tr>
<td>LDAP text (2)</td>
<td>text</td>
<td>delimited</td>
<td>6</td>
<td>1000</td>
</tr>
<tr>
<td>SOAP</td>
<td>text</td>
<td>delimited</td>
<td>6</td>
<td>1000</td>
</tr>
<tr>
<td>Twitter (REST)</td>
<td>text</td>
<td>delimited</td>
<td>6</td>
<td>1825</td>
</tr>
</tbody>
</table>

**TABLE V:** Experiment message trace datasets (available at [56])

We applied our techniques to four real-world protocols, *IMS* [37] (a binary mainframe protocol), *LDAP* [10] (a binary directory service protocol), *SOAP* [38] (a textual protocol, with an Enterprise Resource Planning (ERP) system messaging
system services), and RESTful Twitter [39] (a JSON protocol for the Twitter social media service).

We have used one message trace dataset for each of these protocols. In addition, LDAP has two extra datasets: a dataset with textual representation converted from the binary dataset (denoted by LDAP text (1)), and another textual dataset that was used in our prior work [8] (LDAP text (2)).

We chose these protocols because: (i) they are widely used in enterprise environments; (ii) some of them (LDAP text (2) and SOAP) were used in the evaluation of our prior work [7], [8]; (iii) they represent a good mix of text-based protocols (SOAP and RESTful Twitter) and binary protocols (IMS and LDAP); (iv) they use either fixed length, length encoding or delimiters to structure protocol messages and (v) each of them includes a diverse number of operation types, as indicated by the $Ops$ column. The number of request-response interactions for each test case is shown as column $\#Transactions$ in Table VI.

Our message trace datasets are available for download from our website [36].

### B. 10-fold Cross-Validation Approach and Evaluation Criteria

Cross-validation [40] is a popular model validation method for assessing how accurately a predictive model will perform in practice. For the purpose of our evaluation, we applied the commonly used 10-fold cross-validation approach [61] to all the six case study datasets.

We randomly partitioned each of the original interaction datasets into 10 groups. Of these 10 groups, one group is considered to be the evaluation group for testing our approach, and the remaining 9 groups constitute the training set. This process is then repeated 10 times (the same as the number of groups), so that each of the 10 groups will be used as the evaluation group once.

When running each experiment with each trace dataset, we applied our approach to each request message in the evaluation group, referred to as the incoming request, to generate an emulated response. We recorded the time that our approach took to generate the response for each incoming request. This was used to evaluate the runtime efficiency of our consensus+weighting approach compared to other approaches.

Having generated a response for each incoming request, we utilised five criteria to determine its accuracy, thereby evaluating the ability of our approach to generate protocol-conformant responses. These five criteria are explained as follows using an example shown in Table VI. Consider the incoming request \{id:37, op:A, sn:Durand\} with the associated response \{id:37, op:AddRsp, result:Ok\} in the transaction library. The emulated response is considered to be:

1) **identical** if its character sequence is identical to the recorded (or expected) response (cf. Example (i) in Table VI).

2) **consistent** if it is of the expected operation type and has the critical fields in the payload replicated (cf. Example (ii) in Table VI where id is identical, but some of the other payload differs);

3) **protocol conformant** if its operation type corresponds to the expected response, but it differs in some payload information (cf. Example (iii) in Table VI where both the id and result tags differ);

4) **well-formed** if it is structured correctly (that is, it corresponds to one of the valid response messages), but has the wrong operation type (cf. Example (iv) in Table VI where the generated response is of a valid structure, but its operation type op:SearchRsp does not match the expected operation type op:AddRsp); and

5) **malformed** if it does not meet any of the above criteria (cf. Example (v) in Table VI where operation type op:SearchRsp in the generated response is invalid).

We further consider a generated response to be **valid** if it meets one of the first three criteria, that is, identical, consistent or protocol conformant. Otherwise, a generated response is considered to be invalid.

### C. Compared Techniques

To our knowledge there is no other work that generates application layer responses directly from trace messages. Our comparison is therefore with our prior work, including the Whole Library [7] and the Cluster Centroid [8] approaches. Given an incoming request, the Whole Library approach searches the entire transaction library for its closest matching request to synthesize its response(s). This approach is effective in producing accurate responses. Experimental results revealed that more than 90% of generated responses were valid, as defined by the criteria in Section IV-B. However, it is generally too slow for real-time use.

The Cluster Centroid approach reduces the number of searches to the number of transaction library clusters. Therefore, it can generate responses for real time use, but with less accuracy. We use these two approaches as baselines to evaluate our new technique.

### D. Evaluation Results

1) **Accuracy (RQ1):** The accuracy evaluation is conducted to assess the capability of our approach for generating accurate

<table>
<thead>
<tr>
<th></th>
<th>Expected</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>{id:37,op:AddRsp,result:Ok}</td>
<td>{id:37,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>ii</td>
<td>{id:37,op:AddRsp,result:AlreadyExists}</td>
<td>{id:37,op:AddRsp,result:AlreadyExists}</td>
</tr>
<tr>
<td>iii</td>
<td>{id:15,op:AddRsp,result:Ok}</td>
<td>{id:15,op:AddRsp,result:Ok}</td>
</tr>
<tr>
<td>iv</td>
<td>{id:15,op:SearchRsp,result:Ok,gn:Miao,sn:Diu}</td>
<td>{id:15,op:SearchRsp,result:Ok,gn:Miao,sn:Diu}</td>
</tr>
<tr>
<td>v</td>
<td>{id:37,op:AddRsp,result:Ok}</td>
<td>{id:37,op:AddRsp,result:Ok}</td>
</tr>
</tbody>
</table>

Table VII summarises the evaluation results of Consensus Only, Consensus+Weighting, and our prior work (i.e. Whole Library and Cluster Centroid) experiments for the six test datasets. The Accuracy Ratio column is calculated by dividing the number of valid generated responses by the total number of interactions tested. The last five columns give a more detailed breakdown of the different categories of valid and invalid responses generated.

Table VII shows that the combined Consensus+Weighting approach achieves the highest accuracy overall for the datasets tested. The combined approach achieves 100% accuracy for four of the datasets, and 99.95% and 99.34% for the remaining two (LDAP binary and Twitter, respectively). Twitter is the only case where the Whole Library approach is marginally better.

With respect to the impact for the frequency threshold $f$, the results show that allowing some tolerance (i.e. $f < 1$) in the

Table VII: Evaluation Results of Applying Consensus Sequence Prototype

<table>
<thead>
<tr>
<th>Protocol (binary)</th>
<th>Method</th>
<th>Accuracy Ratio</th>
<th>No.</th>
<th>Valid</th>
<th>Invalid</th>
<th>Identical</th>
<th>Consistent</th>
<th>Conformant</th>
<th>Well-formed</th>
<th>Malformed</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS</td>
<td>Whole Library</td>
<td>75.25%</td>
<td>800</td>
<td>400</td>
<td>202</td>
<td>198</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cluster Centroid</td>
<td>97.88%</td>
<td></td>
<td>400</td>
<td>383</td>
<td>97</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus Only</td>
<td>f=0.5</td>
<td>100%</td>
<td></td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>100%</td>
<td></td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus +</td>
<td>f=0.5</td>
<td>100%</td>
<td></td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weighting</td>
<td>f=1</td>
<td>100%</td>
<td></td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LDAP text (1)</td>
<td>Whole Library</td>
<td>94.12%</td>
<td>2177</td>
<td>248</td>
<td>17</td>
<td>1784</td>
<td>36</td>
<td>92</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cluster Centroid</td>
<td>91.59%</td>
<td></td>
<td>263</td>
<td>17</td>
<td>1714</td>
<td>183</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus Only</td>
<td>f=0.5</td>
<td>87.74%</td>
<td></td>
<td>264</td>
<td>14</td>
<td>1565</td>
<td>334</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>79.97%</td>
<td></td>
<td>259</td>
<td>14</td>
<td>1468</td>
<td>436</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus +</td>
<td>f=0.5</td>
<td>98.77%</td>
<td></td>
<td>278</td>
<td>18</td>
<td>1853</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weighting</td>
<td>f=1</td>
<td>99.95%</td>
<td></td>
<td>278</td>
<td>18</td>
<td>1880</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LDAP text (2)</td>
<td>Whole Library</td>
<td>92.9%</td>
<td>1000</td>
<td>927</td>
<td>2</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cluster Centroid</td>
<td>73.4%</td>
<td></td>
<td>509</td>
<td>225</td>
<td>0</td>
<td>252</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus Only</td>
<td>f=0.5</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=0.8</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus +</td>
<td>f=0.5</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weighting</td>
<td>f=0.8</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>100%</td>
<td></td>
<td>808</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SOAP</td>
<td>Whole Library</td>
<td>100%</td>
<td>1825</td>
<td>150</td>
<td>994</td>
<td>613</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cluster Centroid</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>896</td>
<td>917</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus Only</td>
<td>f=0.5</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=0.8</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consensus+</td>
<td>f=0.5</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weighting</td>
<td>f=0.8</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>f=1</td>
<td>99.34%</td>
<td></td>
<td>0</td>
<td>893</td>
<td>920</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table VII: Evaluation Results of Applying Consensus Sequence Prototype

As illustrated in Equation 1, a pre-defined frequency threshold is required to calculate the consensus sequence prototype. In bioinformatics, a number of investigations have been done for identifying the best threshold [34]. In our experiments, we selected the most popular 3 values, that is, 0.5, 0.8 and 1.

With respect to the impact for the frequency threshold $f$, the results show that allowing some tolerance (i.e. $f < 1$) in the
multiple sequence alignment can yield better results. For the LDAP (binary) dataset the thresholds of $f = 0.5$ and $f = 0.8$ produced significantly higher accuracy than $f = 1$. For the other datasets the threshold had no impact on the accuracy. The general conclusion appears that the results are not very sensitive to the value of the threshold for most scenarios.

An interesting result is that the Consensus+Weighting approach has a higher accuracy than the Whole Library approach, even though the latter uses all the available data points from the trace library (for three datasets Consensus+Weighting is significantly more accurate, for two it has the same accuracy, for one it is slightly lower). The reason for the higher accuracy is that the Consensus+Weighting abstracts away the message payload information sections (using wildcards), so is less susceptible to matching a request to the wrong operation type but with the right payload information, whereas the Whole Library approach is susceptible to this type of error (note the well-formed but invalid responses for the Whole Library approach in Table VII).

The impact of the entropy weightings can only be observed for the LDAP binary dataset. For this test, the weightings significantly improve the accuracy results. For the other datasets, no impact from the weightings can be observed, as the consensus sequence prototype by itself (Consensus Only) already produces 99-100% accuracy.

2) Efficiency (RQ2): Table VIII compares the average response generation time of the Consensus+Weighting approach with the Whole Library and Cluster Centroid approaches. Tests were run on an Intel Xeon E5440 2.83GHz CPUs with 24GB of main memory available. The times represent the average response generation time across all of the requests in the datasets. Table VIII also lists the average response times of the real services from which the original traces were recorded. In order to get a better insight of the runtime performance of our approach, we separately measured matching time and substitution time, results of which are presented in Table VIIIa.

The results show that the Consensus+Weighting approach is very efficient at generating responses, indeed much faster than the real services being emulated. The response generation time is comparable to the Cluster Centroid approach, being faster for some datasets, slower for others. Both of these approaches are about two orders of magnitude faster than the Whole Library approach. However, whereas the Cluster Centroid approach trades off accuracy for speed, the Consensus+Weighting has both high accuracy and speed.

Comparing the matching time versus the substitution time (shown in Table VIIIb) we can observe that the Whole Library approach consumes most of its time during the matching process (because a Needleman-Wunsch alignment is made with every request in the transaction library). Consensus+Weighting and Cluster Centroid have greatly reduced matching times. Twitter has unusually long substitution times, such that for the fast approaches, most time is spent performing the substitution. This is due to the Twitter responses being very long, causing the symmetric field identification (common substring search) to become time consuming.

The Consensus+Weighting generates responses faster than the real services being emulated. This is crucial for supporting testing of an enterprise system under test under realistic performance conditions (delays can be added to slow down the emulated response, but not the other way around). A major limitation of the Whole Library approach is that it cannot generate responses in a time which matches the real services for fast services (such as LDAP).

3) Robustness (RQ3): Our final test is to evaluate whether our Consensus+Prototype approach is robust in generating accurate responses when the clustering process (from Section III-B) is imperfect. For this test we deliberately inject noise into our clusters, i.e. to create clusters where a fraction of the interaction messages are of different operation types. The noise ratios tested were 5%, 10% and 20%. We repeated the experiments with different frequency thresholds (i.e. 0.5, 0.8, and 1).

Table IX summarises the experimental results. The results show that having a frequency threshold below 1 has a very big impact on preserving the accuracy when the clustering is noisy. A threshold of $f = 0.5$ gives the best accuracy. When using this threshold, the accuracy stays above 97% for all datasets, when the noise ratio is 5%. As the noise ratio increases to 20%, the accuracy decreases significantly for binary LDAP, but stays high for the other datasets.

Overall this is a good result. For our approach to work best the clusters produced should be relatively clean, but there is tolerance for a small amount of noise. A noise ratio of 20% is considered very high. Our actual clustering process produced perfect separation (i.e. 0% noise) of interaction messages by operation type for the six datasets tested.
TABLE IX: Response Accuracy for Clusters with Noisy Data

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>Consensus+Weighting</th>
<th>Consensus+Weighting</th>
<th>Consensus+Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f = 0.5</td>
<td>f = 0.8</td>
<td>f = 1</td>
</tr>
<tr>
<td>IMS</td>
<td>99.38%</td>
<td>99.75%</td>
<td>72.25%</td>
</tr>
<tr>
<td>LDAP (1)</td>
<td>97.01%</td>
<td>97.29%</td>
<td>33.49%</td>
</tr>
<tr>
<td>LDAP text (2)</td>
<td>100%</td>
<td>100%</td>
<td>83.74%</td>
</tr>
<tr>
<td>SOAP</td>
<td>98.2%</td>
<td>98.2%</td>
<td>91.3%</td>
</tr>
<tr>
<td>Twitter (REST)</td>
<td>99.34%</td>
<td>99.34%</td>
<td>22.41%</td>
</tr>
</tbody>
</table>

E. Industry Validation

The proposed service emulation technique has been integrated into CA Technologies’ commercial product: CA Service Virtualization. An earlier version of the technique [42] was released as a new feature, named opaque data processing, in version 8.0 of the product and has been sold to customers [43]. Opaque data processing has been used at customer sites to successfully emulate services of protocols not otherwise supported by the product. The emulated protocols included a proprietary extension of IMS and Sun’s ONC/RPC protocol. The present technique is on the backlog for future versions.

F. Threats to Validity

We have identified some threats to validity which should be taken into consideration when generalising our experimental results:

- Our evaluation was performed on six datasets from four protocols. Given the great diversity in message protocols, further testing should be performed on other message protocols.
- The datasets were obtained by randomly generating client requests for services of different protocols. Some real system interactions are likely to be more complicated than those of our datasets. Further testing on real system interactions are warranted.

V. Discussion and Future Work

We have developed an approach for automatically generating service responses from message traces which requires no prior knowledge of message structure or decoders or message schemas. Our approach of using multiple sequence alignment to automatically generate consensus prototypes for the purpose of matching request messages is shown to be accurate, efficient and robust. Wildcards in the prototypes allow the stable and unstable parts of the request messages for the various operation types to be separated. Rather than using the prototypes directly for strict matching (such as using it as a regular expression) we instead calculate matching distance through a modified Needleman-Wunsch alignment algorithm. Since we look for the closest matching prototype, the method is robust even if the prototypes are imperfect. Moreover, this process can match requests which are slightly different to the prototypes, or are of different length to the prototypes. This allows the system to handle requests which are outside of the cases directly observed in the trace recordings. Weighting sections of the prototype with different importance based on the entropy, further improves the matching accuracy.

Our experimental results using the 6 message trace datasets demonstrate that our approach is able to automatically generate accurate responses in real time for most cases. Moreover, we can also see that our approach can also generate accurate responses from imperfect message clusters that contain a small number of messages of different operation types.

One limitation in our current approach is the lack of diversity in the responses generated. We are working on an approach to identify common patterns of all responses in a cluster. A possible solution is to apply multiple sequence alignment to response messages, to distinguish stable positions from variable positions. The variable parts of responses could then be stochastically generated. To further improve the robustness of our approach to noisy clustering, we will utilise outlier detection techniques [44] to remove outliers of clusters before applying the alignment method.

Our approach does not consider the service state history in formulating responses. In practice a stateless model is sufficient in many cases. For example: (i) when the emulation target service is stateless, or (ii) when the testing scenario does not contain two equivalent requests, requiring different state affected responses, or (iii) where the testing scenario does not require highly accurate responses (e.g. performance testing.)

To address this limitation, an avenue of future exploration is to process mine the operation sequences to discover stateful models.

VI. Conclusion

We have developed a new technique for automatically generating realistic response messages from network traces for enterprise system emulation environments that outperforms current approaches. We use the bioinformatics-inspired multiple sequence alignment algorithm to derive message prototypes, adding wildcards for high variability sections of messages. A modified Needleman-Wunsch algorithm is used to calculate message distance and entropy weightings are used in distance calculations for increased accuracy. Our technique is able to automatically separate the payload and structural information in complex enterprise system messages, making it highly robust. We have shown in a set of experiments with four enterprise system messaging protocols a greater than 99% accuracy for the four protocols tested. Additionally they show efficient emulated service response performance enabling scaling within an emulated deployment environment.
REFERENCES


