Efficient Service Substitutions with Behavior-based Similarity Metrics

Joshua Church Amihai Motro
Computer Science Department
Volgenau School of Engineering
George Mason University
Fairfax, VA 22030-4444, USA
Email: jchurch2@gmu.edu

***Abstract*—We define a formal model for information services that incorporates the concept of service similarity. The model places services in metric spaces, and allows for services that have arbitrarily complex inputs and output domains. We then address the challenge of service substitution: finding the services most similar to a given service among a group, possibly large, of candidate services. To solve this nearest neighbor problem efficiently we embed the space of services into a vector space and search for the nearest neighbors in the target space. We report on an extensive experiment that validates both our formalization of similarity and the methods used for finding service substitutions.**

I. Introduction

The increasing deployment of the service-oriented programming paradigm, in which programs are composed by weaving together distributed, platform-independent software components, raises the problem of *service substitution*: How to best replace one software component by another. The substitution may be motivated by a variety of reasons: An existing component might have failed, or there could be alternative services that are of higher quality or lower cost. As the service-oriented paradigm gains popularity, the number of available services increases substantially, making the searches for substitutions more complicated and costly, and raising the need for methods and tools that facilitate such searches [20].

Finding the service (or group of services) that are most similar to a given service can be stated as a k-nearest neighbors problem. As such, it requires establishing a formal notion of *service similarity*. In this paper we describe a formal model for information services that incorporates the concept of service similarity, and we utilize this model for addressing the challenge of finding service substitutions.

Our model has several distinguishing aspects. First, the model considers services as *mappings* of inputs to outputs. These services are memoryless, stateless and free of side-effects. Our view of services is thus quite similar to that advocated by the Representative State Transfer (REST) architectural style [26]. A simple example is a service that returns the current temperature for a given US Zip code. An important feature of the model is that the input and output of a service could be *arbitrarily complex*. For example, a service that returns the value of a stock portfolio given a particular date and a set of stock symbols and their corresponding quantities — has input which is a pair comprising a single value (date) and an arbitrarily-sized set of pairs (stock symbol and quantity). Another feature of the model is that the *semantics* of services need not be represented externally — by means of natural language descriptions, sets of keywords, ontologies, and so on. Instead, when necessary, semantics are inferred from observable behavior.

Our model defines a notion of service *similarity* that is based solely on observable behavior. Similarity is measured with distance *metrics*. Each basic domain (such as real numbers or character strings) is associated with a simple metric, and these metrics are then combined to create complex metrics that could be associated with domains of arbitrary complexity. Given two services with identically structured inputs and outputs, their similarity is derived from the similarity of their outputs for identical inputs. Additionally, the model is extended to deal with service *exceptions*: instances in which services might not provide valid outputs as expected.

A naïve solution to the problem of finding substitutions is to compare the observable behavior of the given service to each of the candidate services. To avoid long computations at the time when the substitutions are needed, similarity measurements should be prepared ahead of time. The cost, however, could be high, because metrics could be complex, and the number of services could be high. To overcome this challenge we describe a method that *embeds* the given set of services into a vector space, where the complexity of measurement is reduced significantly.

Two aspects of this work require experimental validation. The first is the *quality* of the discoveries. It must be demonstrated that our metrics indeed reflect the similarity inherent among services. That is, that services that behave similarly are indeed measured to have strong similarity. Our experiments show the precision of discoveries to be between 88% and 97%. The second aspect is the *performance* of our nearest neighbor methods. Our experiments show that the metric space embedding method reduces search cost by as much as 98%.

Our model is described in two sections: Section III defines services and Section IV defines service similarity. Our approach to the challenge of service substitutions is described in Section V. Section VI details and analyzes the experiments that validates our methods. Finally, Section VII summarizes

978-0-7695-5025-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ICWS.2013.34

the results and sketches future research directions. We begin in Section II with a brief survey of related work.

II. Background

Our work introduces a model for services, and it is therefore helpful to consider other approaches to this subject. The area of service modeling is vast and mostly out of the scope of this work, and we therefore focus on a small but diverse group of formal models for *service discovery*. These are models that examine service descriptions that are informative for search, and address the need to *retrieve* a relevant service.

The most popular approach to service description is the WSDL file. These files describe service operations and parameters as elements in standard XML format. In [5] and [2], custom similarity functions are designed based on information gleaned from a statistical analysis of large WSDL collections. A bag-of-words model is used in [5] to find services that are within some edit distance to indexed WSDL files. In a similar vein, [2] recommends substitutable services based on extended string matching that incorporates tendencies found in WSDL naming conventions. Searching for service alternatives is also our goal; however, our approach builds a model from service *behavior*. In addition, we query our model by *example*; that is, find substitutions using a description of behavior rather than keywords. The use of metrics for defining service similarity is discussed in [14], [8]. The metrics in the former work are based on the similarity of terms that are extracted from the corresponding WSDL descriptions (as well as other related terms). The latter work measures service similarity based on the commonalities in the structures of the services (e.g., the internal processes used). In neither case is the similarity derived from observable behavior.

There are different interpretations of behavior. In [18] and [12], behavior is defined as data flow between service invocations, statements of logic model these connections, and similarity is defined as Boolean combinations of predicates. In [7], graph theory is used to model behavior on the basis of the dependencies among service operations, and similarity is based on graph edit distance. In contradistinction, we represent behavior as sets of input/output observations. This approach is facilitated by instrumenting software as it runs, and then collecting invocation traces in a repository [9]. This type of information can also be gleaned from deployment logs or other histories of program executions [10].

Instead of modeling behavior, [23] presents a service retrieval approach inspired by the relational model. The goal is to optimize QoS queries using techniques from database query processing. The research in [23] assumes sets of functionally equivalent services, whereas our work seeks to *find* equivalent services. In addition, [23] emphasizes differences in QoS parameter when choosing between equivalent services, whereas we emphasize differences in behavior.

An essential concept within our model is the similarity of services. The concept of similarity, it should be pointed out, has been researched extensively in other fields, including psychology [22], [19]. As services are software modules, it is worth mentioning that the notion of similarity is fundamental to many applications in software engineering, including code completion [3], extracting source code snippets [1], visualizing software interfaces [15], and indexing source code repositories [6].

The challenge we address — efficiently finding the services most similar to a given service — is an example of the widely researched nearest neighbor problem [11], [24]. Our criterion for similarity is defined by metric spaces [16], where "nearest to a given service" is interpreted as the service with the smallest distance to the service. Previous applications of metric spaces to service-oriented computing focused on finding geographically close services in ubiquitous computing environments [21], [13]. The goal is to make frequent service lookups efficient by balancing the load among service locators. Similar to our approach, services in [13] are treated as points in a metric space and then embedded into a vector space. However, similarity is not the focus of that work and it assumes that an ontology is available to model service properties.

III. Services

There are different views as to what is a software service. The services considered here are *information services*. An information service is a function that receives data values as input and returns data values as output. As such, an information service is a data object encapsulated in specific input and output protocols. Alternatively, one could consider each service as a small database that is wrapped to process one type of query.

Let A and B be two domains, a service s is therefore a function $s : A \longrightarrow B$.

An example is a service that receives values that are US Zip codes and returns the current temperature in the specified location. Another example is a service that receives a combination of location (latitude and longitude) and returns the location (latitude and longitude) of the three nearest gas stations, along with the names of their oil companies. A third example is a service that receives a date and a set of an arbitrary number of pairs each consisting a stock ticker symbol and the number of shares owned, and returns the total value of the portfolio on that date.

A. Domain Expressions

As the examples illustrate, the domains could have different structures. A domain could be basic (scalar) (e.g., Zip code), it could be a sequence (e.g., latitude and longitude), or it could be a set of sequences (e.g., portfolio). To allow complex input and output we define domains as expressions in which basic domains are combined with two operators: aggregate and sequence.

Given a domain α, the *aggregation* of α, forms a new domain, denoted $\{\alpha\}$, where each element is a *set* of elements of α.

Given domains $\alpha_1, \ldots, \alpha_n$, the *sequencing* of $\alpha_1, \ldots, \alpha_n$, forms a new domain, denoted $(\alpha_1, \ldots, \alpha_n)$, where each element is a *sequence* of n elements in which the ith element is from the domain α_i.

These two operations may be combined to create arbitrarily complex expressions from basic domains. Let α denote a basic domain, then domain expressions Δ are defined with this syntax:

$$\Delta ::= \alpha \mid \{\Delta\} \mid (\Delta, \ldots, \Delta)$$

Denote Z the domain of Zip codes, T the domain of temperatures (e.g., in degrees Fahrenheit), L_1 and L_2 the domains of latitudes and longitudes, respectively, G the domain of oil companies, D the domain of dates, S the domain of stock symbols, and M the domain of money (e.g., US Dollars). Then the three examples of services given earlier have these domains:

$$\begin{aligned} s_1 : Z &\longrightarrow T \\ s_2 : (L_1, L_2) &\longrightarrow \{(L_1, L_2, G)\} \\ s_3 : (D, \{(S, M)\}) &\longrightarrow M \end{aligned}$$

Thus, the output domain of s_2 is a three column table and the input domain of s_3 is a date and a two-column table. Similarly, the expression $\{(\alpha_1, \alpha_2), \{(\alpha_3, \alpha_4, \alpha_5)\}\}$ defines a domain whose values are sets of pairs — the first element of each pair is a pair of scalars, and the second element is a set of triplets of scalars.

B. Signature and Behavior

A service comprises two components, called signature and behavior. Consider a service s with input domain A and output domain B. The *signature* of s is the pair (A, B) of input and output domain expressions. The *behavior* of s is the set of values in the domain A and their matched values in the domain B: $\{(a, s(a)) \mid a \in A\}$. An individual behavior pair $(a, s(a))$ is an *instance* of the service s.[1]

The behavior of a service could be a large set, even infinite. For example, a service that provides the total rainfall in 2012 for each Zip code would have about 43,000 instances. A service that converts degrees Fahrenheit to degrees Celsius would have an infinite number of instances.[2]

IV. Service Similarity

A question that often arises in service-oriented software architectures is whether two given services are "similar" to each other; i.e., whether one service could substitute for the other [4]. To answer this question we must define service similarity. We limit our discussion here to the similarity of services with identical signatures. Hence, the issue is how to define the similarity of two behaviors (of the same signature).

A. Similarity of Domain Elements

To measure the similarity between the elements of a domain we define a metric for that domain. A *metric* for a domain A is a function d from the product $A \times A$ to the real numbers that satisfies

1) $d(a_1, a_2) \geq 0$
2) $d(a_1, a_2) = 0 \iff a_1 = a_2$
3) $d(a_1, a_2) = d(a_2, a_1)$, and
4) $d(a_1, a_2) + d(a_2, a_3) \geq d(a_1, a_3)$

Our approach is to establish distance metrics for the basic (scalar) domains, and then extend these metrics to general domain expressions.[3]

In this paper we consider two basic domains only: real numbers R and character strings S. We note that extending the work to other domains should not be difficult. Distances among real numbers will be measured with the *absolute value metric*, and distances among strings will be measured with the *Levenshtein edit distance*. These two metrics are extended to metrics for general domains as follows.

Let x, y be elements of domain α and assume that $\alpha = (\alpha_1, \ldots, \alpha_n)$; that is, the final domain operation to create α was sequencing. Then x and y are sequences: $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. We define the distance between x and y as the *sum* of the distances between the sequence components:

$$d(x, y) = \sum_{i=1}^{n} d(x_i, y_i) \tag{1}$$

Let x, y be elements of domain α and assume that $\alpha = \{\alpha_1\}$; that is, the final domain operation to create α was aggregation. Then x and y are sets: $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_m\}$. We define the distance between x and y as the Hausdorff distance between the sets. That is, first we calculate the directed distance from x to y: For each element in x we choose the smallest of its distances to the m elements of y, and then we choose the largest of the n distances thus obtained; we similarly calculate the directed distance from y to x; and finally we adopt the highest of the two distances:

$$d(x, y) = \max\{\max_{x_i \in x}\{\min_{y_j \in y}\{d(x_i, y_j)\}\}, \max_{y_j \in y}\{\min_{x_i \in x}\{d(y_j, x_i)\}\}\} \tag{2}$$

As the domains α_i used to create α could themselves be complex, these definitions are reapplied recursively, until all distances are calculated in basic domains, using the two basic metrics.

To illustrate, assume a domain $\alpha = (S, \{R\})$ and two values from this domain: ('Jack', $\{1, 3, 5\}$) and ('Jill', $\{5, 8, 11, 14\}$). The Levenshtein distance between 'Jack' and 'Jill' is 3; the Hausdorff distance between $\{1, 3, 5\}$ and $\{5, 8, 11, 14\}$ is 9; and the overall distance between ('Jack', $\{1, 3, 5\}$) and ('Jill', $\{5, 8, 11, 14\}$) is therefore $3 + 9 = 12$.

We note that the eventual distance function on α thus defined is indeed a metric. This follows from (1) the distances defined on the basic domains (absolute value and Levenshtein) are both metrics, (2) the distance between sequences as defined in Equation 1 is a metric,[4] and (3) the distance between

[1] This distinction between signature and behavior is similar to the distinction between intension and extension, common in logic and in databases.

[2] For now we ignore the fact that this type of service is better stored as a formula, such as $C = (F - 32) \cdot 5/9$.

[3] Note that metrics measure distance, and eventually we shall convert distance to similarity.

[4] This metric is termed the *product metric*, as the new space is a product of the component spaces.

aggregates as defined in Equation 2 is a metric.[5]

Finally, we define the *similarity* between two elements of a domain as the *reciprocal* of the distance:

$$sim(x_1, x_2) = 1/d(x_1, x_2) \tag{3}$$

B. Similarity of Behavior

Recall that we interpreted the similarity of two services (of the same signature) as the similarity of their behaviors. Assume two services with identical signatures:

$$\begin{aligned} s_1 : A &\longrightarrow B \\ s_2 : A &\longrightarrow B \end{aligned}$$

The similarity between s_1 and s_2 is defined as the *average* of the similarity of their outputs for all possible inputs. Let $|A| = n$. Then:

$$sim(s_1, s_2) = \frac{1}{n} \sum_{a \in A} sim(s_1(a), s_2(a)) \tag{4}$$

As an example, assume a small domain $A = \{1, 2, 3, 4\}$, a service s_1 that multiplies its input by 5, a service s_2 that multiples its input by 6, and a service s_3 that adds 10 to its input. Their behavior tables can be combined:

A	s_1	s_2	s_3
1	5	6	11
2	10	12	12
3	15	18	13
4	20	24	14

To calculate the similarity of any two services we average the similarity of their outputs for the four possible inputs: $sim(s_1, s_2) = 1/2.5 = 0.4$, $sim(s_1, s_3) = 1/4 = 0.25$ and $sim(s_2, s_3) = 1/5 = 0.2$.

C. Service Exceptions

Similar to null values in database tables, behavior tables may have missing values as well, where the service cannot deliver an output value for a particular value of the input. These missing values are termed *service exceptions*. Consider this small example in which service exceptions are denoted with —:

A	s_1	s_2
1	b_1	c_1
2	b_2	—
3	—	c_3
4	b_4	c_4
5	b_5	—
6	b_6	c_6
7	—	—

For calculating service similarity in the presence of exceptions, our approach is to substitute each distance that cannot be calculated due to an exception with the *minimal* known similarity. That is,

$$\begin{aligned} sim(s_1, s_2) = \frac{1}{7} \cdot (&sim(b_1, c_1) + sim(b_4, c_4) + sim(b_6, c_6) \\ &+ 4 \cdot \min\{sim(b_1, c_1), sim(b_4, c_4), sim(b_6, c_6)\}) \end{aligned}$$

Denote A_0 the subset of A in which both services do not have exceptions. Let $|A| = n$ and $|A_0| = n_0$. Therefore the number of instances in which at least one service has an exception is $n - n_0$. Then

$$\begin{aligned} sim(s_1, s_2) = \frac{1}{n} \cdot (&\sum_{a \in A_0} sim(s_1(a), s_2(a)) \\ &+ (n - n_0) \cdot \min_{a \in A_0} \{sim(s_1(a), s_2(a))\}) \end{aligned} \tag{5}$$

Obviously, in the absence of exceptions, Equation 5 reduces to Equation 4.

V. Service Substitutions

In Sections III–IV we defined a model for services with service similarity. We now consider the following problem: Given a service s_0 and a group of services S, find the service in S that is most similar to s_0. This *nearest neighbor* problem arises frequently in service-oriented software architectures, when one of the services in the architecture fails and needs to be replaced by the most similar service from a directory of available services. Ideally, this substitution (often called *healing*) would be done "automatically" (i.e., without programmer's intervention) using the most similar service. In practice, however, it is safer to allow the programmer to choose from several substitution alternatives, which is a *k-nearest neighbor* problem. Before we describe our approach to the problem, we discuss two pragmatic adjustments to our model.

A. Reducing the Dimension of Behavior Tables

The definitions of our model assumed that behaviors are available in their entirety, and to determine the similarity of two behaviors we must know their entire sets of instances. In practice, this assumption is impractical for a variety of reasons. First, in many instances services are stored as executable code, not as tables. Second, even when stored in tables, these tables are "wrapped" in protocols that permit queries that access few instances at a time. Finally, even when the tables are available, their sizes may be substantial thus requiring costly calculations.

A practical approach is to obtain *samples* of behavior: Typically, a sample of size n is obtained by a sequence of n single-instance queries to the service. In [4] we showed how relatively small samples (of about 16 instances) may be obtained that convey satisfactorily the "traits" of a behavior. In the following discussion we shall assume behaviors that are *samples*.

[5]The final step in that process, taking the maximum of the two directed distances, assures the required symmetry.

B. Directed Similarity

Our samples are obtained by individual requests to the service, and when a service request results in an exception, we enter an "exception" into the behavior table. However, keeping in mind that our main application is service substitution, our calculation of service similarity (behavior similarity) would no longer be symmetric.

When considering a substitution of s_1 with s_2, the similarity calculation will ignore all instances in which s_1 has an exception in its behavior; and when considering a substitution of s_2 with s_1, our calculation will ignore all instances in which s_2 has an exception in its behavior. The reason is that we assume that the service to be replaced is operating correctly, and its exceptions were a result of illegal out-of-range requests. We expect the "new" service to deliver valid output only in the instances in which the "old" service delivered valid output.

Denote A_1 the subset of A in which s_1 does not have exceptions and denote A_2 the subset of A in which s_2 does not have exceptions. Recalling that A_0 is the subset for which both services do not have exceptions, we have $A_1 \cap A_2 = A_0$.

Let $|A_1| = n_1$ and $|A_2| = n_2$. Therefore the number of instances in which s_2 has exception but s_1 does not is $n_1 - n_0$. The *directed similarity* from s_1 to s_2 is

$$\vec{sim}(s_1, s_2) = \frac{1}{n_1} \cdot \Big(\sum_{a \in A_0} sim(s_1(a), s_2(a)) + (n_1 - n_0) \cdot \min_{a \in A_0} \{sim(s_1(a), s_2(a))\} \Big) \quad (6)$$

Observe that $\vec{sim}(s_1, s_2)$ in Equation 6 is derived from $sim(s_1, s_2)$ in Equation 5, by substituting n_1 for n in two places.

In the previous example, when s_2 is considered a substitute for s_1 the directed similarity would be:

$$\vec{sim}(s_1, s_2) = \frac{1}{5} \cdot (sim(b_1, c_1) + sim(b_4, c_4) + sim(b_6, c_6) + 2 \cdot \min\{sim(b_1, c_1), sim(b_4, c_4), sim(b_6, c_6)\})$$

whereas when s_1 is considered a substitute for s_2 the directed similarity would be:

$$\vec{sim}(s_2, s_1) = \frac{1}{4} \cdot (sim(b_1, c_1) + sim(b_4, c_4) + sim(b_6, c_6) + \min\{sim(b_1, c_1), sim(b_4, c_4), sim(b_6, c_6)\})$$

We begin with a naïve treatment of this *k*-nearest neighbor problem, and we then offer a more efficient solution.

C. Exhaustive Search

As mentioned earlier, the optimal service substitution problem assumes a given service s_0 and a group of services S, and finds the k services in S that are most similar to s_0; that is, services $s_1, \ldots, s_k \in S$, such that for $1 \leq i < k - 1$ $sim(s_0, s_i) \geq sim(s_0, s_{i+1})$ and for every other service $s \in S$ $sim(s_0, s) \leq sim(s, s_k)$.

The simplest solution, of course, is to conduct an exhaustive search. That is, calculate the similarity between s_0 and each service in the group S, while at each point maintaining a list of the top k services discovered so far.

Assume $|S| = n$. In practice, we pre-compute all n^2 similarities and store them in an $n \times n$ matrix. When it is necessary to substitute a service s_j, the j'th row in the matrix is retrieved and top k values are extracted. The positions of these values point to the best substitutions. This "brute-force" algorithm has complexity $O(n^2)$. Note, however, that calculating each of the similarities, especially for services over complex domains, could become costly. Our next method attempts to reduce this cost.

D. Metric Space Embedding

A common method to reduce the effort of calculating distances (or similarities) between elements in a given metric space, is to embed the elements in another metric space, so that the distances between the mapped elements are similar to the original distances, but are (hopefully) considerably simpler to calculate [17]. Formally, let ψ be an embedding of the metric space (A, d) into the metric space (A', d'). We want to find an embedding ψ such that for any two elements $a_1, a_2 \in A$:

$$d(a_1, a_2) \approx d'(\psi(a_1), \psi(a_2))$$

where $\approx$ denotes approximation.

Let (A, d) be a metric space. We extend the metric d to measure the distance between an element $a \in A$ and a subset $X \subseteq A$ as follows:

$$d(a, X) = \min_{x \in X} \{d(a, x)\}$$

That is, the distance between an element and a set is the minimum distance between the element and any element in the set. Now let $X_1, X_2, \ldots, X_m$ be subsets of A. We map every element a to a *vector* of its distances to these subsets:

$$\psi : a \longrightarrow (d(a, X_1), d(a, X_2), \ldots, d(a, X_m))$$

Such a mapping is called a *Lipschitz embedding*, and $X_1, X_2, \ldots, X_m$ are its *reference sets*. The intuition is that the distance between two elements a_1 and a_2 will be captured adequately by the distance between their embedded vectors:

$$\begin{aligned} d(a_1, a_2) &\approx d'(\psi(a_1), \psi(a_2)) \\ &= d'((d(a_1, X_1), d(a_1, X_2), \ldots, d(a_1, X_m)), \\ &\quad (d(a_2, X_1), d(a_2, X_2), \ldots, d(a_2, X_m))) \end{aligned}$$

A particular version of a Lipschitz embedding commonly used in situations similar to ours selects references that are *singleton* sets. Typical metrics for the target vector space are either the Chebyshev metric, that measures the distance between vectors as the maximum distance between corresponding coordinates, or any of the L_p metrics (of which the Chebychev metric is a limit).

Denote the reference points $p_1, p_2, \ldots, p_m$. Then with the Chebyshev metric we have

$$\begin{aligned} d(a_1, a_2) &\approx d'(\psi(a_1), \psi(a_2)) \\ &= \max_{i=1,\ldots,m} \{|d(a_1, p_i) - d(a_2, p_i)|\} \end{aligned}$$

And with the Euclidean metric (L_2):

$$\begin{aligned} d(a_1, a_2) &\approx d'(\psi(a_1), \psi(a_2)) \\ &= \sqrt{\sum_{i=1,\ldots,m} (d(a_1, p_i) - d(a_2, p_i))^2} \end{aligned}$$

In practice, after choosing the m reference points (also called *pivots*), we pre-compute the $n \cdot m$ distances between each of the n services and each of the m pivots. These distances are then used to calculate all service-service similarities, which are stored in an $n \times n$ matrix as before. The new cost is now $O(m \cdot n)$. Assuming $m < n$, it is an improvement over $O(n^2)$.

VI. Experimentation

Two aspects of the work described so far require validation. First, it must be demonstrated that the method for constructing similarity measures for domains of arbitrary complexity, described in Section IV, generates measures of good quality. Then, it must be demonstrated that the k-nearest neighbor method described in Section V arrives at high quality results at low computational costs. This section describes experiments that seek to validate both these aspects.

The first challenge was to design an experiment that will determine whether our similarity measures are successful in conveying the similarity *inherent* in a set of services. That is to say, the set of services has an inherent *structure*, and the issue is whether our similarity measure can detect this structure. We translated this problem to a problem of *classification*: Assume a set S of n services, and a grouping of the services into m distinct groups, where the services in each group are constructed to be inherently similar to each other. Now, given a service $s \in S$, our similarity measure was used to find the k services in S that are most similar to s. The set of k discovered services was compared to the group to which s belongs, using the measure of *precision*: the ratio of services that are indeed in the group of s to the total number k of services discovered.

Two independent experiments were conducted, each with a different output signature. In the first experiment the services had an output signature of the type $(S, \{R\})$; that is, the output of each service was a character string followed by a set of numbers (which is similar to the example shown in Section IV). In the second experiment the services had an output signature of the type $(R, \{S\}, \{S\})$; that is, the output of each service was a triplet: a number followed by two set of strings (for example, a service that receives the GPS coordinates of a location and returns the Zip code of that location plus a set of movie theaters and a set of restaurants in that area).

To imbue services in the same group with similar behaviors, each group of services was randomly assigned the parameters of a distribution, and then each service in the group randomly chose values for its behavior instances from that distribution. For example, assume the output includes a set of numbers. The parameters of the distribution are two ranges: a range from which the cardinality of the set is chosen, and a range from which the values of the elements of the set are chosen. Thereafter, for each service in the group, a cardinality is randomly chosen from the first range and a corresponding set of values is randomly chosen from the second range.

The parameters used in both experiments were as follows. The set S consisted of n=1,000 randomly generated services, where each service was described in a behavior sample of 8 instances.[6] Then, m=10 groups of equal size were generated, with each group consisting of 100 services. The $n \times n$ similarity matrix was then computed, and a total of 278 queries were attempted (this number was chosen to assure statistical significance that is higher than 0.95). The precision of each search was calculated with neighborhood sizes of k = 1, 3, 5, and 10. For the embedding, we used 10 randomly chosen pivot services (this choice is justified later). The $n \times n$ matrix was constructed as described at the end of Section V, and the same 278 queries were attempted and scored. To assure statistical significance, this entire experiment was repeated 100 times.

A. Quality

Figure 1 plots the results of finding k-nearest neighbors in both the original space of services and in the target vector space. The horizontal axis is the size of k, and the vertical axis is the overall precision. Two plots reflect first experiment and two plots reflect the second experiment.

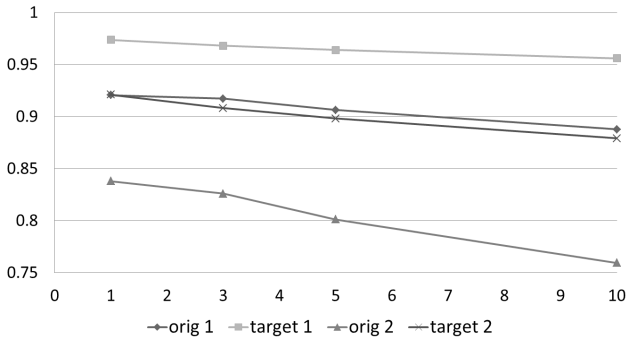

Fig. 1. Precision at 4 neighborhood sizes

Several interesting observations are due. First, quite surprisingly, in both experiments, the performance in the target space was significantly better than the performance in the original space. Whereas we were hoping that the embedding will improve performance (performance is discussed next) while not harming precision too much — in effect, precision improved by 5–6 percentage points. Second, across all levels of k this precision (in the target space) is kept at the impressive range of 88%–97% — a definite validation of our methods. Third, performance in the second experiment was clearly lower than in the first experiment. This can be attributed to the fact the second signature involves more string comparisons than the first signature. The services discovered and presented to the programmer bear similarity to search engine results. Good information retrieval algorithms give higher precision in earlier

[6] As mentioned in Section V, tables of this magnitude have been shown to be effective samples.

pages. The fourth observation is that the decline in precision with the increase in k is of a similar nature. Indeed, it is evidence that our methods percolate the better solutions to the top. Finally, a *significance* test of the precision showed that these results could not happen by chance.

Overall, it can be concluded safely that our methods — both the design of the signature metric and the approach to discovering the best substitute services – are highly effective.

B. Performance Gain

The improvement in quality after embedding the space of services in a vector space was significant and surprising. Nonetheless, recall that the main reason for this transformation was to improve performance. We expected performance to improve because (1) fewer similarities would be calculated, and (2) the calculation of each similarity would be simpler. For validation, we measured the time (minutes of CPU time) that was required for a nearest neighbor search in the original space and in the target space. This time included 10 runs, each requiring completing the $n \times n$ similarity matrix, and, for each of 278 queries, finding the 10 most similar services. We then formed this ratio:

$$Performance\ Gain = \frac{Time\ in\ original\ space}{Time\ in\ target\ space}$$

The first experiment yielded a ratio of $83.17/1.43 = 58.16$ (98.3% reduction), and the second experiment yielded a ratio of $287.00/3.29 = 87.23$ (98.8% reduction). Without a doubt, such gains in performance, coupled with the previously discussed improvement in precision, testifies to the advantage of using metric space embedding.

C. Number of Pivots

As observed at the end of Section V, the gain in performance achieved by the embedding method is in reducing n^2 to $n{\cdot}m$. It is therefore beneficial to keep the number of reference points m low. We tested 8 different numbers of pivots: 1, 5, 10, 30, 100, 200, 500 and 1,000 (recall that the total number of services is 1,000). Figure 2 plots the precision in each experiments for all but the largest 3 numbers (which were essentially the same as for 100). Precision was measured for the case $k = 1$ (i.e., search for the most similar service).

As can be observed, in both experiments, precision exceeds 0.92 when only 10 pivots are used, and it exceeds 0.96 when 30 pivots are used (it is almost flat thereafter). In other words, excellent results were achieved with an embedding that uses just 1–3% of the services as its reference points. This accounts for the large performance gains observed earlier. Again, results for the first experiment were slightly better.

D. Exceptions

Finally, recall that our model allows for the possibility of exceptions in service behavior. Therefore, it is important to validate the robustness of our methods in the presence of exceptions. Figure 3 plots, for each experiment, the search performance in both the original space of services and the target vector space when exceptions are introduced into service behavior.

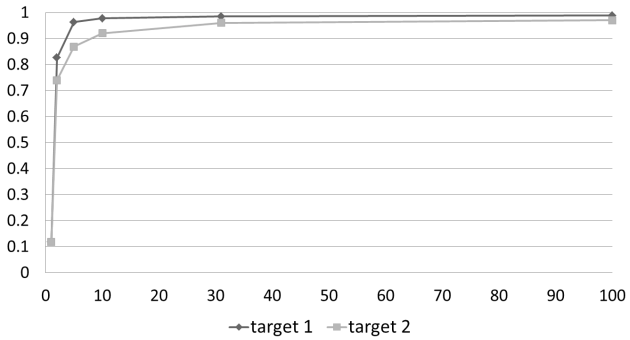

Fig. 2. Precision at different number of pivots

The experiment was done with a neighborhood of $k = 1$ with exception percentages of 1%, 5%, 10% and 20%. For comparison, the plots also include precision when there are no exceptions (0%). As before, performance in the target vector space was consistently higher, and performance in the first experiment was noticeably better. Observe that at exception rates under 10% precision is kept above 0.79. As one is unlikely to deploy services with exception rates higher that 10%, these results are quite encouraging.

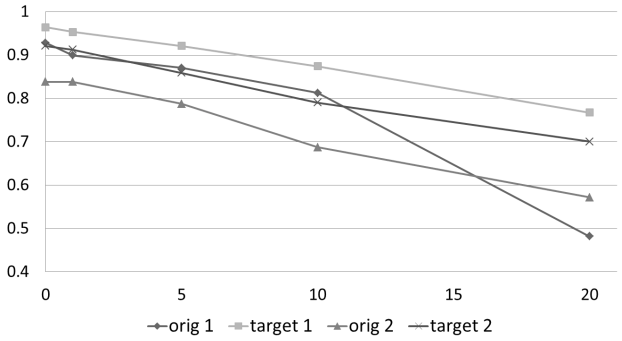

Fig. 3. Precision at 4 percentages of service exception

VII. Conclusion

The increased reliance on publicly available services for creating *ad hoc*, low-cost applications brought about an urgent need for locating service substitutes. As services are usually discovered in large repositories, this problem is naturally stated in terms of a *k*-nearest neighbors problem, which, in turn, requires a formal method for measuring service similarity. In our opinion, it is more productive to gauge similarity on the basis of the intrinsic behavior of services than on meta-information that is associated with services.

In this paper we defined a formal, yet pragmatic method of measuring similarity, and we demonstrated that services discovered on the basis of this measure are indeed, with very high level of precision, similar to the subject of the search

(even in the presence of service exceptions). Additionally, we showed how the discovery of k-nearest neighbors can be approximated in a vector space created by Lipschitz embeddings, with enormous gains in performance.

The results have been encouraging and suggest that expanding the scope of this work may offer additional rewards. We suggest here several such opportunities.

The use of Lipschitz embeddings with a rather small number of pivots brought about a large gain in performance. One research direction is to examine further methods for improving the performance of searches; for example, using *vantage-point trees* to locate the most similar services.

Our experimentation used services that were synthesized. While this allowed us to scale up searches to a very large number of services, it would be desirable to apply our solutions in an environment of actual services. An environment that we plan to explore is biomedical services [25].

Practically the only limiting assumption made in the model was that similarity is defined only for services that share the same signature; that is, services with identically-structured inputs and outputs. Thus, we cannot assess the similarity of two weather services when one provides the temperature at a given location, whereas the other provides both the temperature and the barometric pressure. Expanding our treatment to include similarity between services with different, though overlapping, signatures would be beneficial, as it could yield substitutes that (when tweaked slightly) would be even more satisfactory.

The model and methods that we developed can be applied to other problems in the area of service-oriented programming. For example, our methods for discovery and ranking can be adapted to provide a tool that *recommends* services based on examples. As another example, the similarity measures could be used to interconnect services in a semantic network, allowing users to *explore* large networks of services.

REFERENCES

[1] Sushil K. Bajracharya, Joel Ossher, and Cristina V. Lopes. Leveraging usage similarity for effective retrieval of examples in code repositories. In *Proceedings of the 18th ACM SIGSOFT international Symposium on Foundations of Software Engineering*, pages 157–166. ACM, 2010.

[2] M. Brian Blake and Michael F. Nowlan. A web service recommender system using enhanced syntactical matching. In *Proceedings of ICWS 2007, IEEE International Conference on Web Services*, pages 575–582. IEEE, 2007.

[3] Marcel Bruch, Martin Monperrus, and Mira Mezini. Learning from examples to improve code completion systems. In *Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering*, pages 213–222. ACM, 2009.

[4] Joshua Church and Amihai Motro. Learning service behavior with progressive testing. In *Proceedings of SOCA 11, IEEE International Conference on Service-Oriented Computing and Applications*, pages 1–8. IEEE, 2011.

[5] Xin Dong, Alon Halevy, Jayant Madhavan, Ema Nemes, and Jun Zhang. Similarity search for web services. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pages 372–383. VLDB Endowment, 2004.

[6] Mark Grechanik, Chen Fu, Qing Xie, Collin McMillan, Denys Poshyvanyk, and Chad Cumby. A search engine for finding highly relevant applications. In *Proceedings of ICSE 2010, ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pages 475–484. IEEE, 2010.

[7] Daniela Grigori, Juan Carlos Corrales, and Mokrane Bouzeghoub. Behavioral matchmaking for service retrieval. In *Proceedings of ICWS 2006, IEEE International Conference on Web Services*, pages 145–152. IEEE, 2006.

[8] Akın Günay and Pınar Yolum. Structural and semantic similarity metrics for web service matchmaking. In *E-Commerce and Web Technologies, LNCS 4655*, pages 129–138. Springer, 2007.

[9] Murali Haran, Alan Karr, Alessandro Orso, Adam Porter, and Ashish Sanil. Applying classification techniques to remotely-collected program execution data. *ACM SIGSOFT Software Engineering Notes*, 30(5):146–155, 2005.

[10] Ahmed E. Hassan. The road ahead for mining software repositories. In *FoSM 2008, Frontiers of Software Maintenance*, pages 48–57. IEEE, 2008.

[11] Gisli R. Hjaltason and Hanan Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):530–549, 2003.

[12] Martin Junghans, Sudhir Agarwal, and Rudi Studer. Behavior classes for specification and search of complex services and processes. In *Proceedings of ICWS 2012, IEEE International Conference on Web Services*, pages 343–350. IEEE, 2012.

[13] Saehoon Kang, Daewoong Kim, Younghee Lee, Soon J. Hyun, Dongman Lee, and Ben Lee. A semantic service discovery network for large-scale ubiquitous computing environments. *ETRI journal*, 29(5):545–558, 2007.

[14] Fangfang Liu, Yuliang Shi, Jie Yu, Tianhong Wang, and Jingzhe Wu. Measuring similarity of web services based on WSDL. In *Proceedings of ICWS 10, International Conference on Web Services*, pages 155–162. IEEE, 2010.

[15] Collin McMillan, Mark Grechanik, Denys Poshyvanyk, Qing Xie, and Chen Fu. Portfolio: finding relevant functions and their usage. In *Proceedings of the 33rd International Conference on Software Engineering*, pages 111–120. ACM, 2011.

[16] Mícheál O'Searcoid. *Metric spaces*. Springer, 2006.

[17] Hanan Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.

[18] Zhongnan Shen and Jianwen Su. Web service discovery based on behavior signatures. In *Services Computing, 2005 IEEE International Conference on*, volume 1, pages 279–286. IEEE, 2005.

[19] Roger N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[20] Jonathan Sillito, Gail C. Murphy, and Kris De Volder. Asking and answering questions during a programming change task. *IEEE Transactions on Software Engineering*, 34(4):434–451, 2008.

[21] Liying Tang and Mark Crovella. Virtual landmarks for the internet. In *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pages 143–152. ACM, 2003.

[22] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[23] Qi Yu and Manjeet Rege. A relational approach for efficient service selection. In *Proceedings of ICWS 2009, IEEE International Conference on Web Services*, pages 719–726. IEEE, 2009.

[24] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity search*. Springer, 2006.

[25] Jia Zhang, Ravi K. Madduri, Wei Tan, Kevin Deichl, John Alexander, and Ian T. Foster. Toward semantics empowered biomedical web services. In *Proceedings of ICWS 2011, IEEE International Conference on Web Services*, pages 371–378. IEEE, 2011.

[26] Haibo Zhao and Prashant Doshi. Towards automated RESTful web services composition. In *Proceedings of ICWS 2009, IEEE International Conference on Web Services*, pages 189–196. IEEE, 2009.