Top-k web services compositions: A fuzzy-set-based approach
Karim Benouaret, Djamal Benslimane, Allel Hadjali

To cite this version:
Karim Benouaret, Djamal Benslimane, Allel Hadjali. Top-k web services compositions: A fuzzy-set-based approach. ACM - Symp. on Applied Computing (SAC), Mar 2011, Taiwan. pp.1038-1043, 2011. <hal-00670704>

HAL Id: hal-00670704
https://hal.archives-ouvertes.fr/hal-00670704
Submitted on 15 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Top-k Service Compositions: A Fuzzy Set-Based Approach

Karim Benouaret  
Lyon 1 University  
69622, Villeurbanne, France  
karim.benouaret@liris.cnrs.fr

Djamal Benslimane  
Lyon 1 University  
69622, Villeurbanne, France  
djamal.benslimane@liris.cnrs.fr

Allel Hadjali  
Enssat-University of Rennes 1  
22305, Lannion, France  
allel.hadjali@enssat.fr

ABSTRACT
Data as a Service (DaaS) is a flexible way that allows enterprises to expose their data. Composition of DaaS services provides bridges to answer queries. User preferences are becoming increasingly important to personalizing the composition process. In this paper, we propose an approach to compose DaaS services in the context of preference queries where preferences are modeled by means of fuzzy sets that allow for a large variety of flexible terms such as 'cheap', 'affordable' and 'fairly expensive'. The proposed approach is based on RDF-based query rewritings to take into account the partial matching between individual DaaS services and parts of the user query. Matching degrees between DaaS services and fuzzy preference constraints are computed by means of different constraints inclusion methods. Such degrees express to which extent a service is relevant to the resolution of the query. A fuzzification of Pareto dominance is also proposed to better rank composite services by computing the score of services. The resulting scores are then used to compute the top-k DaaS service compositions that cover the user query.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Query formulation, Search process, Selection process; H.3.5 [Online Information Services]: Web-based services

General Terms
Measurement, Theory, Performance

Keywords
Fuzzy dominance, Web service, Top-k compositions, fuzzy preferences queries.

1. INTRODUCTION
Nowadays, companies are increasingly moving towards a service-oriented architecture for data sharing on the Web by putting their data sources behind Web services, thereby providing a well-documented and interoperable method of interacting with their data [4, 5, 7]. In particular, if no single Web service can satisfy the functionality required by the user, there should be a possibility to combine existing services together in order to fulfill the request. In this context, we talk about DaaS (Data as a Service) Service Composition, where services correspond to calls, i.e. queries, over the data sources.

When searching for items to purchase over the Web, customer preferences are critical in the search. A more general and suitable approach to model preferences is based on fuzzy sets theory [6]. Fuzzy sets are very well suited to the interpretation of linguistic terms, which constitute a convenient way for a user to express her/his preferences. For example, when expressing preferences about the 'price' of a car, users often employ fuzzy terms like 'rather cheap', 'affordable' and 'not expensive'.

However as DaaS services and providers proliferate, a large number of candidate compositions that would use different services may be used to answer the same query. Hence, it is important to set up an effective service composition framework that would identify and retrieve the most relevant services and return the top-k compositions according to the user preferences.

The following example illustrates a typical scenario related to our previous discussion, showing the challenges in finding the top-k service compositions.

Running example. Let us consider an e-commerce system supporting users to buy cars. The system has access to the set of services described in Table 1. The symbols '$' and '?' denote inputs and outputs of DaaS services, respectively. Services providing the same functionality belong to the same service class. For instance, \{S_{21}, S_{22}, S_{23}, S_{24}\} belong to the same class S_2. Each service has its (fuzzy) constraints on the data it manipulates. For instance, the cars returned by S_{21} are of cheap price and short warranty.

Assume that a user Bob wants to buy a car. He sets his preferences and submits the following query Q_1: 'return the French cars, preferably at an affordable price with a warranty around 18 months and having a normal power with a medium consumption'. Bob will have to invoke S_{11} to retrieve the French automakers, he can then invoke one or more of the services S_{21}, S_{22}, S_{23}, S_{24} to retrieve the French cars along with their prices and warranties, finally he will invoke one or more of the services S_{31}, S_{32}, S_{33}, S_{34} to retrieve the power and the consumption of retrieved cars. This manual process is painful and tedious. It raises the follow-
2. SERVICE COMPOSITION-BASED PREFERENCE QUERIES ANSWERING

Table 1: Example of DaaS Services

<table>
<thead>
<tr>
<th>Service</th>
<th>Functionality</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{11}(x, y)$</td>
<td>Returns the automakers $y$ whose country is $x$</td>
<td>-</td>
</tr>
<tr>
<td>$S_{21}(x, y, z, t)$</td>
<td>Returns the cars $y$ along with their prices $z$ and warranties $t$ for a given automaker $x$</td>
<td>$z$ is cheap, $t$ is short</td>
</tr>
<tr>
<td>$S_{22}(x, y, z, t)$</td>
<td>-</td>
<td>$z$ is accessible, $t$ is long</td>
</tr>
<tr>
<td>$S_{23}(x, y, z, t)$</td>
<td>-</td>
<td>$z$ is expensive, $t$ is long</td>
</tr>
<tr>
<td>$S_{24}(x, y, z, t)$</td>
<td>-</td>
<td>$z$ is [9000, 14000], $t$ is [6, 24]</td>
</tr>
<tr>
<td>$S_{31}(x, y, z)$</td>
<td>Returns the power $y$ and the consumption $z$ for a given car $x$</td>
<td>$y$ is weak, $z$ is small</td>
</tr>
<tr>
<td>$S_{32}(x, y, z)$</td>
<td>-</td>
<td>$y$ is ordinary, $z$ is approximately 4</td>
</tr>
<tr>
<td>$S_{33}(x, y, z)$</td>
<td>-</td>
<td>$y$ is powerful, $z$ is high</td>
</tr>
<tr>
<td>$S_{34}(x, y, z)$</td>
<td>-</td>
<td>$x$ is [60, 110], $z$ is [3.5, 5.5]</td>
</tr>
</tbody>
</table>

In our approach, user preferences are modeled using fuzzy sets [6]. Formally, a fuzzy set $F$ on a referential $X$ is characterized by a membership function $\mu_F : X \rightarrow [0, 1]$, where $\mu_F(x)$ represents the grade of membership of $x$ in $F$. In particular, $\mu_F(x) = 1$ reflects full membership of $x$ in $F$, $\mu_F(x) = 0$ absolute non-membership, and $0 < \mu_F(x) < 1$ partial membership. Namely having $x, y \in F$, $x$ is more preferable than $y$ iff $\mu_F(x) > \mu_F(y)$. If $\mu_F(x) = \mu_F(y)$, $x$ and $y$ are equally preferred. In practice, membership functions are often of trapezoidal form represented by the quadruplet $(A, B, a, b)$. A regular interval $[A, B]$ can be seen as a fuzzy set represented by the quadruplet $(A, B, 0, 0)$. The semantics of all fuzzy terms of Table 1 are available in the URL: http://vm.liris.cnrs.fr:36880/FuzzyTerms/.

2.2 DaaS services

Let us assume that DaaS services are partitioned into different classes of services. A class $S_j$ represents services with the same inputs, outputs, and providing the same functionality but different (fuzzy) constraints. A DaaS service $S_{ji}$ of class $S_j$ is described as a predicate $S_{ji}(x_j, y_j) : < G_{ji}(x_j, y_j, z_j), C_{ji} >$ where:

- $X_j$ and $Y_j$ are the sets of input and output variables of $S_{ji}$, respectively. Input and output variables are also called distinguished variables. They are prefixed with the symbols ‘$’ and ‘?’, respectively.
- $G_{ji}(x_j, y_j, z_j)$ represents the functionality of the service. This functionality is described as a semantic relationship between input and output variables. $Z_j$ is the set of existential variables relating $X_j$ and $Y_j$.
- $C_{ji} = \{C_{j1}, ..., C_{j8}\}$ is a set of constraints expressed as intervals or fuzzy sets on $X_j$, $Y_j$ or $Z_j$ variables.

$X_j$ and $Y_j$ variables are defined in the WSDL description of DaaS services. Functionality $G_j$ and constraints $C_{ji}$ of a DaaS service $S_{ji}$ are added to the standard WSDL descriptions in the form of annotations. The annotations are represented in the form of SPARQL queries. For instance, the following SPARQL query illustrates the functionality and service constraints of the DaaS service $S_{2i}$:

```
SELECT ?y ?z ?t
?C rdf:type Car ?C hasName ?y ?C hasPrice ?x
?C hasWarranty ?t ?x is 'cheap' ?t is 'short'}
```

2.3 Discovering Relevant Services

Let $Q$ be a conjunctive preference query. We use an efficient RDF query rewriting algorithm to discover the parts
we are interested in is how to rank DaaS services in each class \( S_j \) to select the most relevant services and how to rank generated DaaS service compositions to select the top-\( k \) ones that answer the fuzzy query \( Q \).

### 3. FUZZY DOMINANCE AND FUZZY SCORES

#### 3.1 Fuzzy dominance v.s. Pareto dominance

Services of the same class \( S_j \) have the same functionality, they only differ in terms of constraints, providing thus different matching degrees. Individual matching degrees of services could be aggregated. One method is to assign weights to individual degrees and, for instance, compute a weighted average of degrees. In so doing, users may not know enough to make trade-offs between different relevancies using numbers (average degrees). Users thus lose the flexibility to select their desired answers by themselves. Computing the skyline [3] is as a natural solution to overcome this limitation. The skyline consists of the set of points which are not dominated by any other point.

**Definition 1.** (Pareto dominance) Let \( u \) and \( v \) be two d-dimensional points. We say that \( u \) dominates \( v \), denoted by \( u \succ v \), if \( \forall i \in [1,d] \), \( u_i \geq v_i \land \exists k \in [1,d] \), \( u_k > v_k \).

Pareto dominance is not always significant to rank-order points that, however, seem comparable from a user point of view. To illustrate this situation, let \( u = (u_1, u_2) = (1, 0) \) and \( v = (v_1, v_2) = (0.90, 1) \) be two matching degrees. In Pareto order, we have neither \( u \succ v \) nor \( v \succ u \), i.e. the instances \( u \) and \( v \) are incomparable. However, one can consider that \( v \) is better than \( u \) since \( v_2 = 1 \) is too much higher than \( u_2 = 0 \), contrariwise \( v_1 = 0.90 \) is almost close to \( u_1 = 1 \). This is why it is interesting to fuzzify the dominance relationship to express the extent to which a matching degree (more or less) dominates another one. In line with the general fuzzification dominance approach discussed in [8], we define below a fuzzy dominance relationship that relies on particular membership function of a graded inequality of the type strongly larger than.

**Definition 2.** (Fuzzy dominance) Given two d-dimensional points \( u \) and \( v \), we define the fuzzy dominance to express the extent to which \( u \) dominates \( v \) such as:

$$\deg(u \succ v) = \sum_{i=1}^{d} \mu_\succ(u_i, v_i)$$

Where \( \mu_\succ(u_i, v_i) \) expresses the extent to which \( u_i \) is more or less (strongly) greater than \( v_i \). \( \mu_\succ \) can be seen as a monotonic membership function defined as:

$$\mu_\succ(x, y) = \begin{cases} 
1 & \text{if } x - y > \lambda + \varepsilon \\
0 & \text{if } x - y \leq \varepsilon \\
\frac{x - y}{\lambda} & \text{otherwise}
\end{cases}$$

Where \( \lambda > 0 \), i.e. \( \succ \) is more demanding than the idea of 'strictly greater'. We should also have \( \varepsilon \geq 0 \) in order to ensure that \( \succ \) is a relation that agrees with the idea of 'greater' in the usual sense.

Let us reconsider the previous instances \( u = (1, 0) \), \( v = (0.90, 1) \), with \( \varepsilon = 0 \) and \( \lambda = 0.2 \). We have \( \deg(u \succ v) = 0.25 \) and \( \deg(v \succ u) = 0.5 \). This is more significant than \( |u \succ v| = |v \succ u| = 0 \) provided by Pareto dominance, where \( |u \succ v| = 1 \) if \( u \succ v \), 0 otherwise. In the following sections, we use the defined fuzzy dominance to compute scores of services and compositions.
3.2 Associating fuzzy score with a service

It is well known that under a single matching degree method (mono criteria), the dominance relationship is unambiguous. When multiple CIM are applied (multi-criteria), resulting in different matching degrees for the same couple of constraints, the dominance relationship becomes uncertain. The model proposed in [10], namely probabilistic skyline overcomes this problem. Contrariwise, Skoutas et al. show in [11] the limitations of the probabilistic skyline to rank Web services and introduce the Pareto dominating score of individual services. We generalize this score to fuzzy dominance and propose a fuzzy dominating score (FDS). An FDS of a service \( S_{ji} \) indicates the average extent to which \( S_{ji} \) dominates the whole services of its class \( S_j \).

Definition 3. (Fuzzy dominating score of a service) The fuzzy dominating score (FDS) of a service \( S_{ji} \) in its class \( S_j \) is defined as:

\[
FDS(S_{ji}) = \frac{1}{(|S_j| - 1)|M|} \sum_{h=1}^{|M|} \sum_{k \in S_j \cap C} \sum_{r=1}^{|M|} \text{deg}(S^h_{ji} > S^r_{jk}) \quad (3)
\]

where \( S^h_{ji} \) is the matching degree of the service \( S_{ji} \) obtained by applying the \( h \)th CIM. The term \((|S_j| - 1)\) is used to normalize the FDS score and make it in the range \([0...1]\).

Table 3 shows the fuzzy dominating scores of the DaaS services of our running example.

3.3 Associating fuzzy score with a composition

Different DaaS service compositions can be generated from different \( S_j \) service classes to answer a user query. To rank such generated compositions, we extend the previous FDS definition to service composition and associate each composition with an FDS. The FDS of a composition \( C \) is an aggregation of different FDSs of its component services.

Let \( C = \{ S_{j1}, ..., S_{jn} \} \) be a composition of \( n \) services. Let also \( d = d_1 + ... + d_n \) be the number of user preference constraints where \( d_i \) is the number of constraints involved in the service \( S_{ji} \). The FDS of \( C \) is then computed as follows:

\[
FDS(C) = \frac{1}{d} \sum_{j=1}^n d_j \cdot FDS(S_{ji}) \quad (4)
\]

It is important to note that not all compositions are valid. A composition \( C \) of services is valid if (i) it covers the user query \( Q \), (ii) it contains one and only one service from each service class \( S_j \), and (iii) it is executable. A composition is said to be executable if all input parameters necessary for the invocation of its component services are bound.

4. TOP-K DAAS SERVICE COMPOSITIONS

4.1 An Efficient Generation of Top-k Compositions

A straightforward method to find the top-\( k \) compositions that answer a query is to generate all possible compositions, compute their scores, and return the top-\( k \) ones. However, this approach results in a high computational cost, as it needs to generate all possible compositions, whereas, most of them are not in the top-\( k \). In the following, we provide an optimization technique to find the top-\( k \) compositions. This technique allows eliminating relevant services \( S_{ji} \) from their classes \( S_j \) before generating the compositions, i.e. services that we are sure that if they are composed with other ones, the obtained compositions are not in the top-\( k \). The idea is: we first compute the score of each service in its class, then only the best services in each class are retained, after that we compose the retained services, finally, we compute the score of the obtained compositions and return the top-\( k \) ones. To this end, we introduce the following theorem\(^1\).

THEOREM 1. Let \( C = \{ S_{j1}, ..., S_{jn} \} \) be a composition. Let top-\( k \)\( (S_j) \) (resp. top-\( k \)\( (\text{compositions}) \)) be the top-\( k \) services of the class \( S_j \) (resp. the top-\( k \) compositions). Then, \( \exists S_{ji} \in C; S_{ji} \notin \text{top-}k(S_j) \Rightarrow C \notin \text{top-}k(\text{compositions}) \)

This theorem means that the top-\( k \) sets of the different service classes \( S_j \) are sufficient to compute the top-\( k \) compositions that answer the query at hand.

The fourth column of Table 3 shows the top-\( k \) (where \( k = 2 \)) services in each service class using the FDS scores. Thus, relevant DaaS services that are not in the top-\( k \) of their classes are eliminated. They are cross out in Table 3. The other DaaS services are retained. The top-\( k \) DaaS service compositions are generated from the different top-\( k \) \( S_j \) classes. Table 4 shows the possible compositions along with their fuzzy scores and the top-\( k \) compositions of our running example.

Table 3: Services’ scores and top-k services

<table>
<thead>
<tr>
<th>Services</th>
<th>Class</th>
<th>Score</th>
<th>Top-k</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{j1} )</td>
<td>( S_1 )</td>
<td>0.487</td>
<td>( S_{i1} )</td>
</tr>
<tr>
<td>( S_{j2} )</td>
<td>( S_2 )</td>
<td>0.653</td>
<td>( S_{i2} )</td>
</tr>
<tr>
<td>( S_{j3} )</td>
<td>( S_3 )</td>
<td>0.593</td>
<td>( S_{i3} )</td>
</tr>
<tr>
<td>( S_{j4} )</td>
<td>( S_4 )</td>
<td>0.743</td>
<td>( S_{i4} )</td>
</tr>
</tbody>
</table>

Table 4: Compositions’ scores and top-k ones

<table>
<thead>
<tr>
<th>Compositions</th>
<th>Score</th>
<th>Top-k</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 = { S_{j11}, S_{j22}, S_{j32} } )</td>
<td>0.623</td>
<td>( C_2 )</td>
</tr>
<tr>
<td>( C_2 = { S_{j13}, S_{j22}, S_{j34} } )</td>
<td>0.698</td>
<td>( C_4 )</td>
</tr>
<tr>
<td>( C_3 = { S_{j11}, S_{j24}, S_{j32} } )</td>
<td>0.566</td>
<td>( C_4 )</td>
</tr>
<tr>
<td>( C_4 = { S_{j11}, S_{j24}, S_{j34} } )</td>
<td>0.640</td>
<td>( C_4 )</td>
</tr>
</tbody>
</table>

4.2 Top-k DaaS Service Compositions Algorithm

The algorithm, hereafter referred as \( TKDSC \), computes the top-\( k \) DaaS service compositions according to the fuzzy scores. The algorithm proceeds as following.

Step.1 compute the matching degrees (lines 1-13). Each class whose services cover a query component is added to the list of relevant classes. If its services touch the query’s user preferences, we compute its different matching degrees according to the number of methods.

\(^1\)Proofs excluded due to lack of space
Algorithm 1 TKDSC

Require: $Q$ a preference query, $S$ a set of service classes, $M = \{m_1, ..., m_{|M|}\}$ a set of methods, $k \in \mathbb{N}$, $\varepsilon, \lambda$

1: for all $S_j$ in $S$ do
2: $S \leftarrow \text{random}(S_j, 1)$;
3: if $\exists q_j \in Q$; $\text{cover}(S_j, q_j)$ then
4: $R \leftarrow R \cup S_j$;
5: if $P_{q_j} \neq \emptyset$ then
6: for all $S_{ji}$ in $S_j$ do
7: for all $m$ in $M$ do
8: MatchingDegree($C_{ji}, P_{q_j}, m$);
9: end for
10: end for
11: end if
12: end if
13: end for
14: for all $S_j$ in $R$ do
15: if $P_{q_j} = \emptyset$ then
16: $\text{top-k}, S_j \leftarrow \text{random}(S_j, k)$;
17: else
18: for all $S_{ji}$ in $S_j$ do
19: ServiceScore($S_{ji}$);
20: end for
21: $\text{top-k}, S_j \leftarrow \text{top}(k), S_j$;
22: end if
23: end for
24: $C \leftarrow \text{ComposeServices}(\text{top-k}, S_{j1}, ..., \text{top-k}, S_{jm})$;
25: for all $C$ in $C$ do
26: CompositionScore($C$);
27: end for
28: return $\text{top}(k), C$;
ceived much attention in the service computing community during the last years. Taking user preferences into account allows to rank candidate services/compositions and return only the best k ones to the user. Hereafter, we review some works for ranking and selecting Web services. Authors in [9] propose a method to rank semantic web services. It is based on computing the matching degree between a set of requested NFPs (Non-Functional Properties) and a set of NFPs offered by the discovered Web services. NFPs cover QoS aspects, but also other business-related properties such as pricing and insurance. Semantic annotations are used for describing NFPs and the ranking process is achieved using some automatic reasoning techniques that exploit the annotations. Wang et al. [12] propose a system for conducting qualitative Web service selection in the presence of incomplete or conflicting user preferences. The paradigm of CP-nets is used to model user preferences. The system utilizes the history information (stored in the system) of users to amend the preferences of active users, thus improving the results of service selection. In [13], the authors present a new ranking method, called ServiceRank. It considers the QoS aspects as well as the social perspectives of services (such as how they invoke each other via service composition). A service that provides good QoS and is invoked more frequently by others is more trusted by the community and will be assigned a higher rank.

The works the most related to our proposal are those proposed in [11, 14]. In [11], Skoutas et al. address ranking and clustering Web services and propose methods based on the notion of dominance; these methods apply multiple matching criteria without aggregating the match scores of individual services parameters. The notion of Web service dominance introduced relies on the one of uncertain dominance discussed in [10]. Two distinct scores are defined for assessing when a service is objectively interesting: (i) the dominated score of a service to indicate the average number of services by which it is dominated (ii) the dominating score of a service to indicate the average number of services that it dominates. Algorithms for computing the top-k Web services according to each score are provided. To achieve a balance between dominated and dominating scores, a third ranking score for a service that combines both, is introduced.

In [14], Yu and Bouguettaya address the problem of uncertainty inherent in the Quality of Web Service (QoWS) and compute the skylines from service providers (referred to as service skylines). A service skyline can be regarded as a set of service providers that are not dominated by others in terms of QoWS attributes that interest all users. To this end, a concept called p-dominant service skyline is defined. A provider S belongs to the p-dominant skyline if the probability that S is dominated by any other provider is less than p. The authors provide also a discussion about the interest of p-dominant skyline w.r.t. the notion of p-skyline proposed in [10]. They show also that the former, on one hand, avoids requiring from users to transform personal preferences into numerical weights and, on the other hand, will not miss any ‘really good’ providers (i.e., the ones that provide consistently high QoWS).

7. CONCLUSIONS

In this paper, we addressed the top-k service compositions problem to answer fuzzy preference queries. The key notion of our approach is the fuzzy dominance relationship introduced to measure the extent to which a vector of matching degrees more or less dominates another one. This kind of fuzzy dominance is then used to both rank DaaS services in their classes and DaaS compositions in order to compute the top-k service compositions. Although we have focused on the DaaS services to answer a query, the proposed approach could be applied in other cases to compute, for instance, the top-k QoS-based service compositions.

8. REFERENCES