On the Evaluation of Semantic Web Service Matchmaking Systems

Vassileios Tsetsos, Christos Anagnostopoulos, Stathes Hadjiefthymiades
Pervasive Computing Research Group, Communication Networks Laboratory,
Dept of Informatics & Telecommunications, University of Athens, Panepistimiopolis, Athens
15784, Greece, Tel: +30 210 7275362
{b.tsetsos, bleu, shadj}@di.uoa.gr

Abstract

Semantic Web Services are generally considered as the evolution of conventional Web Services. Semantic information included in the service descriptions enables the development of advanced matchmaking schemes, capable of assigning degrees of match to the discovered services. In this paper, we address issues related to the evaluation of the retrieval effectiveness of semantic matchmaking systems. Our main position is that conventional evaluation schemes do not fully capture the added value of service semantics nor do they take into account the assigned degrees of match, supported by the majority of discovery engines. Through some preliminary experiments, we show that a generalization of the evaluation process based on fuzzy set theory techniques can lead to more accurate and meaningful evaluation results.

1. Introduction

Service discovery and selection is one of the central topics in distributed systems research. The topic gained additional popularity with the introduction of Web Services (WS), which enable XML-based interactions between applications. WS discovery is performed with the aid of the UDDI registries that support keyword-based matching between the textual descriptions of the user request and the published/advertised services. However, according to the Semantic Web vision, WS will eventually be replaced by Semantic Web Services (SWS). SWS are, essentially, a metadata layer that allows for more expressive description of service capabilities, used both for service advertisements (formed by the service providers) and requests (formed by the service requestors). Such metadata is represented through semantic Web technologies like ontologies and rules. The SWS ecosystem involves, apart from the aforementioned description facilities, more sophisticated (mainly logic-based) matchmaking algorithms [17]. Such algorithms are used in order to improve the discovery process, since only the advertisements that are “logically relevant” to the user request are retrieved.

Typical information that the SWS matchmaking exploits is: service Inputs, Outputs, Preconditions and Effects (a.k.a. IOPE attributes). However, no matter on which elements of a service description the matching algorithm is applied to, the most important problem in matchmaking is that it is unrealistic to expect relevant advertisements and requests to be in perfect match. Hence, adopting a hard-decision approach, would result in ignoring services that partially match a service request. The problem is aggravated if we take into account that the service request may not fully capture the requestor’s intention. In that case, one can observe the paradox of a service advertisement partially matching the issued service request and perfectly matching the requestor’s intention (or the inverse situation). Moreover, the output obtained in response to a query is not ranked in terms of importance to the user; thus, each retrieved item is assumed to be as important as any other. This would further hinder the application of relevance feedback techniques.

The concept of the “Degree of Match” (DoM) was introduced for dealing with these problems. DoM can be informally defined as a value from an ordered set of values that expresses how similar two entities are, with respect to some similarity/relevance metric(s). In the field of SWS such entities may be services, IOPE attributes or specific service operations. Hence, a service matchmaking algorithm calculating the DoM can be used for ranking the discovered services according to their relevance to the issued request. Several, slightly different, variants of DoM have been proposed in the relevant literature [1][2][3]. For instance, in [2], the possible (logic-based) DoM between a service request R and an advertisement S are shown in Table 1.

Table 1. Degrees of Match (as defined in [2])

DoM	Definition (informal)
EXACT	If the inputs and outputs of R are equivalent concepts with the inputs and outputs of S, respectively
PLUGIN	If the outputs of S are direct subclasses of the outputs of R and the inputs of R are subsumed by the inputs of S in the domain ontology
SUBSUMES	If the outputs of S are subsumed by the outputs of R and the inputs of R are subsumed by the inputs of S in the domain ontology
SUBSUMED-BY	If the outputs of R are direct subclasses of the outputs of S and the inputs of R are subsumed by the inputs of S in the domain ontology[1]
FAIL	If none of the above logic-based criteria holds true (this DoM is termed "logic-based fail" in [2])

Similarly to other retrieval systems, such as Web search engines, SWS discovery systems should be evaluated in terms of *performance* and *retrieval effectiveness*. Throughout this paper, the term "performance" implies the computational complexity, response times, etc. of the system. On the other hand, "retrieval effectiveness" (or simply "effectiveness") illustrates how good the system is in discovering *relevant* services, as they have been specified by a domain expert.

Many researchers have undertaken extensive performance assessment efforts for measuring retrieval times and the scalability of the available tools. However, what is still missing from the current literature is a quantitative analysis and comparison of the retrieval effectiveness of the discussed approaches. To the best of our knowledge, only a few researchers have contacted such experimental evaluations, as discussed in Section 5. There are several reasons for this situation, with the main being the lack of established evaluation metrics, methodologies and service test collections. Most evaluation efforts apply well-known Information Retrieval (IR) metrics to SWS discovery evaluation. The most popular are precision and recall, along with their combined metrics, e.g., F-measure [7]. Such metrics have been widely used in other fields where matchmaking is applied (e.g., schema matching [4]). However, unless in-depth analysis is undertaken, one cannot be sure that these metrics apply in the same way to service discovery. Such analysis should define, for instance, how can the various service discovery objectives be expressed through precision/recall metrics and how do they relate to the concepts of false positives and false negatives.

In this paper, we address some issues on the evaluation of the retrieval effectiveness of SWS matchmaking systems. Our main thesis is that traditional evaluation schemes neither fully capture the added value of service semantics nor do they take into account the service ranking (expressed through DoM), supported by the majority of SWS discovery engines. The main objective of this work is to propose a revised evaluation scheme for SWS discovery based on sound IR theories. Such scheme is able to improve the evaluation tools by making their results more fine-grained.

The rest of the paper is organized as follows. In Section 2, we describe in more detail the Boolean evaluation adopted by most researchers and indicate its shortcomings for SWS matchmaking. In Section 3, we present a generalized retrieval evaluation scheme, based on [5], and discuss its application in the domain of SWS discovery. Additionally, we perform a preliminary qualitative comparison between Boolean and generalized fuzzy evaluation. In Section 4, we provide some directions for dealing with certain practical issues raised by such evaluation approach. In Section 5, we briefly present the evaluation metrics and methods used by other related works. The paper concludes with some suggestions for future research.

2. Service Retrieval Evaluation Schemes

2.1 Evaluation Basics

Figure 2.a depicts a reference scheme for service retrieval evaluation. The involved entities are the service advertisements *(S_i)* published in a service registry, the service request *R* posed by the user, and the matchmaking engine that is responsible for the actual service discovery. In essence, the matchmaking engine assigns a Degree of Match ***e(R, S_i)*** to every service advertisement S_i. In IR terminology, such value is called RSV (Retrieval Status Value). These values determine the ranking of the final advertisements for a specific request *R*. In order to evaluate the matchmaking engine effectiveness some expert mappings *r(R, S_i)* (i.e., relevance assessments between *R* and S_i) should be available/pre-specified (see Figure 2.b). Hence, the vectors ***r*** and ***e*** are defined as:

$$\boldsymbol{r}: Q\times S\rightarrow W, \quad \boldsymbol{e}: Q\times S\rightarrow W$$

where Q is the set of all possible service requests, S the set of service advertisements and W the set of values denoting the degree of relevance (for ***r***) or degree of

[1] the original definition is somewhat different since it also includes a similarity-based condition

match (for ***e***) between a request from Q and a service from S. Both ***r*** and ***e*** may assume various types of values: Boolean (W={0,1}), real numbers (W=[0,1]), fuzzy terms (W={"irrelevant", "relevant", …}), etc. Given these informal definitions, *the evaluation of a matchmaking engine is the determination of how closely vector **e** (delivered by the engine) approximates vector **r** (specified by domain experts).*

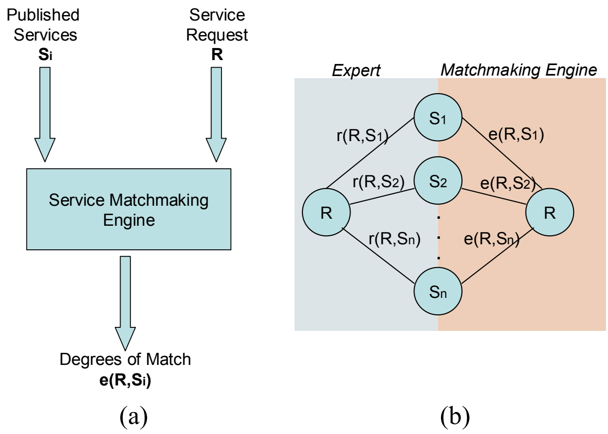


Figure 2. Service Retrieval Evaluation. a) reference scheme, b) relevance assessments

Table 2. Evaluation schemes

Evaluation Scheme	RSVs – e(R,S_i)	Expert Mappings – r(R,S_i)
EVS1	Boolean	Boolean
EVS2	Multi-valued	Multi-valued

Depending on whether W is a Boolean set or not, we may end up with having two different evaluation schemes, which are shown in Table 2 and will be further discussed in the rest of the paper. Note that we assume that the request is always a Boolean conjunctive query, i.e., no special weights are assigned to the various query elements such as inputs and outputs.

The EVS1 (Boolean) scheme is the traditional scheme used in the relevant literature (see Section 5) and is described in the following subsection. EVS2, on the other hand, is proposed as a more appropriate alternative for SWS discovery evaluation, since it overcomes the shortcomings of the former scheme.

2.2 Boolean Evaluation Scheme

2.2.1 Description

Such scheme (referred to as EVS1 in Table 2) is based on Boolean request-to-advertisement matching functions. Whenever a requestor issues a request, the system computes, for each service advertisement, the corresponding RSV (i.e., membership function *e* shown in Eq. 1). Such value indicates whether or not the advertisement is "relevant" to the request.

$$\boldsymbol{e}: \mathrm{Q} \times \mathrm{S} \rightarrow \{0,1\} \qquad \text{Eq. (1)}$$

with, $\boldsymbol{e}(R, S_i) = 1$ if advertisement $S_i \in$ S is "relevant" to $R \in$Q and $\boldsymbol{e}(R, S_i) = 0$ if S_i is "irrelevant" to R. Analogously, the domain experts assign the values "1" and "0" to all advertisements of the service collection with respect to the request *R*.

In this case, standard measures such as precision and recall [7] are used for measuring the system performance. Recall, R_B, is defined as the percentage of the number of "retrieved and relevant" advertisements over the number of "relevant" advertisements in the service collection. Precision, P_B, is defined as the percentage of the number of "retrieved and relevant" advertisements over the number of "retrieved" advertisements. Eq.2 shows the definitions for these metrics where RT and RL are the sets of "retrieved" and "relevant" advertisements, respectively.

$$R_B = \frac{|RT \cap RL|}{|RL|}, \; P_B = \frac{|RT \cap RL|}{|RT|} \qquad \text{Eq. (2)}$$

2.2.2 Shortcomings

As already mentioned, the majority of SWS matchmaking systems support multiple DoM values. However, in the case of Boolean evaluation, such information is not taken into account and, hence, *the evaluation does not fully exploit the available service semantics*. Specifically, if we have to evaluate a matching engine supporting four degrees of match, we should divide the whole range of service rankings to two complementary sets through a threshold value (i.e., minimum acceptable DoM). We assume that all the discovered services having a DoM above the threshold are "relevant" and all other services "irrelevant" to the service request. In other words we quantize the multi-valued ***e*** to a Boolean ***e′***. This is depicted in Figure 3, where the value "A" represents a perfect match and "D" denotes a non-match.

Another shortcoming of the Boolean evaluation scheme comes from the perspective of the expert mappings. A Boolean relevance assessment for service advertisements is too coarse-grained, which is in

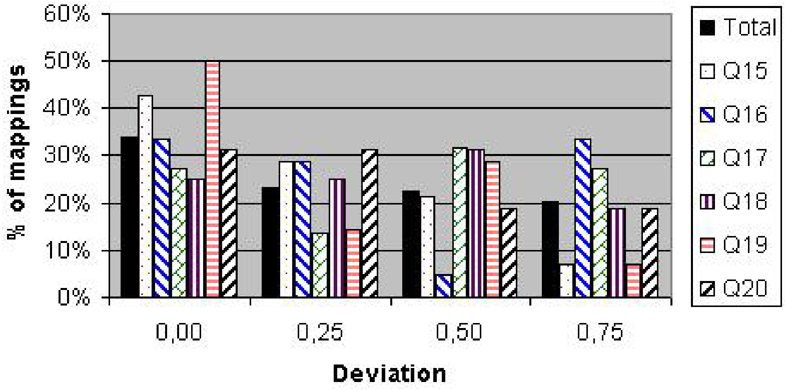
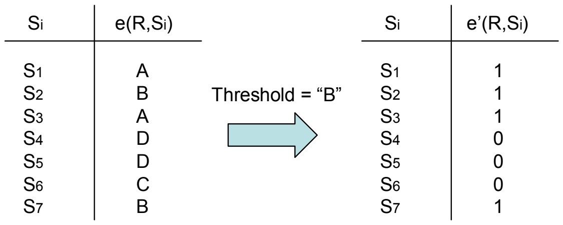


Figure 4. Mapping deviation between Boolean and multi-valued relevance assessment



Si	e(R,Si)
S1	A
S2	B
S3	A
S4	D
S5	D
S6	C
S7	B

Threshold = "B"

Si	e'(R,Si)
S1	1
S2	1
S3	1
S4	0
S5	0
S6	0
S7	1

Figure 3. "Booleanization" of multi-valued matches. The threshold filters out any semantics assigned to the matched services

contrast to one of the objectives of SWS discovery, i.e., more effective and accurate service retrieval. To further illustrate this we have performed some multi-valued expert mappings for the "education" subset of the TC2 service collection [9] and we have computed their deviation from the existing Boolean assessments of TC2. This service subset contains 135 service advertisements, 6 queries (Q15-Q20) and the relevance service set for each query (see also the TC2 manual available in [9]). The indicative expert mappings we have performed assumed the values W={0, 0.25, 0.5, 0.75, 1} for the degree of relevance (in fact, the expert used the linguistic terms shown in Figure 5 but here we consider only the value representing the defuzzyfied membership curve of each separate term). The deviation of such relevance assessment from the Boolean case is the difference *1-X*, where "1" stands for the Boolean relevance and $X \in W$. The total deviation and the per-query deviation is depicted in Figure 4. From this analysis we observe that, on average, only about 1/3 of the Boolean assessments fully agree with the multi-valued ones (i.e., deviation = 0). Moreover, only a total of about 60% of the Boolean mappings are sufficiently close (i.e., deviation ≤ 0.25) to the non-Boolean ones. In other words, the "Boolean domain expert" may assign full relevance between a request and an advertisement even if she observes only partial or minor relevance (i.e., cases where deviation ≥ 0.5).

To summarize, we observe the following problems with a Boolean evaluation scheme:

a) the multi-valued matchmaking algorithm execution results are transformed to Boolean values, and, thus, the matchmaking and service semantics is ignored
b) such transformation involves the definition of a threshold. The assignment of an optimal value to this system parameter is not a trivial task.
c) the Boolean relevance assessments are too coarse-grained and do not always reflect the real intention of the domain expert

Given these shortcomings, EVS1 cannot accurately assess how close the discovered services are to the truly relevant services. We propose that a solution to these problems would be to use the EVS2 scheme of Table 2, as discussed in the following section.

3. A Generalized Fuzzy Evaluation Scheme for Service Retrieval

In order to deal with the problems identified above, we assume that a service discovery system is a generalized retrieval system, similar to that described in [5]. In such a system the evaluation is performed according to the EVS2 scheme (see Table 2). Hence, the following changes are implied for the evaluation of the discovery effectiveness:

a) expert mappings are performed in a non-Boolean manner (in our case *W* is a set of fuzzy linguistic terms),
b) The degrees of match (RSVs) supported by the matchmaking engine are represented by non-Boolean values (we have used fuzzy terms similar to the expert mapping terms), and

c) The standard Boolean precision and recall measures (see Eq. 2), are substituted by generalized evaluation measures.

These changes are further discussed in the following subsections.

3.1 Fuzzy Relevance Assessment by Domain Experts

Evaluating how relevant two service descriptions are (i.e., the request and an advertisement) is a very difficult and context-dependent task. Among the factors that affect the discussed relevance assessment are the characteristics of the service description language (e.g., its expressiveness), whether or not users have previous experience with the specific services. Hence, due to the multifaceted nature of this task an expert should be able to make assessments more fine-grained than those supported by a Boolean scheme. One way to do this would be to use numeric weights in the specification of the expert mappings. However, according to L. Zadeh, the concept of "Relevance" can be characterized as "amorphic" [15], i.e., its main characteristic is that we cannot define it mathematically due to its complexity. In other words, the use of such weights would force the user to quantify a set of qualitative and rather vague concepts (i.e., to quantify the impact of the abovementioned factors). Moreover, when using numeric weights one should be well aware of and define their semantics [6]. Hence, when attempting to qualify phenomena related to human perception (like expert mappings between a set of queries and relevant services) it is very helpful to use words in natural language (i.e., linguistic terms) instead of numerical values.

Hence, we propose the adoption of a fuzzy[2] linguistic approach in order to discriminate the services by their relevance to the request. Such fuzzy linguistic approach associates a linguistic descriptor to each service advertisement, such as "somewhat relevant" or "very relevant", instead of numerical values. The theoretical basis of such approach is Fuzzy Set Theory (FST) [12], which has been used in order to achieve a mathematical formulation of the use of weights for handling uncertain information in various representation levels. Specifically, linguistic values are modeled by means of fuzzy linguistic variables [13]. Several IR systems have been proposed that adopt a fuzzy linguistic approach to model either weights in the query or membership functions during query evaluation. In our case, we use an ordinal fuzzy linguistic approach, which defines the linguistic term set by means of an ordered list of linguistic terms with respect to a fuzzy linguistic variable. A simplified definition of a linguistic variable is:

Definition. A *linguistic variable* is characterized by a tuple(*L, H(L)*). *L* is the linguistic name of the variable (e.g., "relevance") and *H(L)* denotes the linguistic term set of *L*, i.e., the set of names of linguistic values of *L* (e.g., "irrelevant", "somewhat relevant"). Each linguistic value can be denoted by a fuzzy variable *u* ranging across a universe of discourse *U.* The membership degree of an element $u \in U$ is defined by a membership function μ_u, such that: μ_u:U→[0,1]. A value of 0 means no membership, whilst a value of 1 indicates full membership.■

(A more complete definition can be found in [6]).

In the context of the SWS evaluation, the name of the linguistic variable *L* used by the domain experts is "relevance" and the set *H(L)* is defined as: *H*("relevance") = {"irrelevant", "slightly relevant", "somewhat relevant", "relevant", "very relevant"}. For instance, if a service request $R \in Q$ is "slightly relevant" to a service advertisement $S_i \in S$, then $\mu_{slightly\_relevant} \simeq 1$. The membership functions of such terms may be either evenly distributed or not in the interval [0, 1] with respect to the ordered structure of the corresponding linguistic terms [14]. We have assumed linear trapezoidal membership functions for capturing the vagueness of the various linguistic terms. Figure 5 depicts the membership functions of the aforementioned terms.

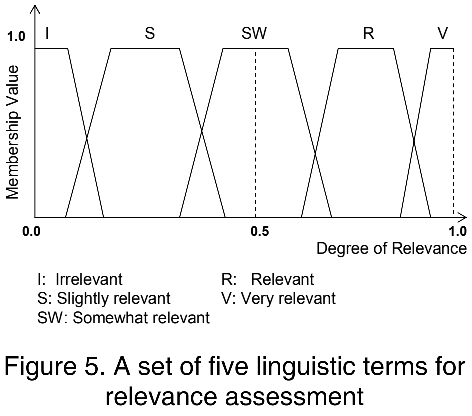


Figure 5. A set of five linguistic terms for relevance assessment

[2] It should be noticed that the adjective "fuzzy" in the context of this paper does not refer to fuzzyness in the matchmaking process per se (e.g., fuzzy queries or fuzzy matchmaking algorithm) but only to the way of modeling relevance and degrees of match, i.e., through fuzzy variables.

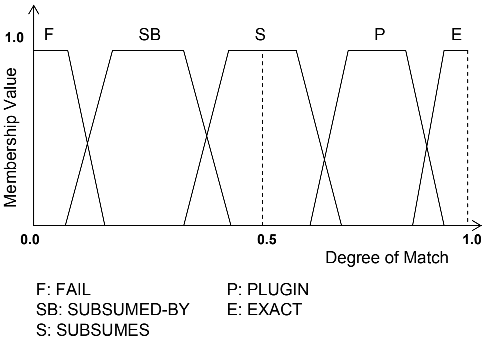


Figure 6. Fuzzy degrees of match

3.2 Fuzzification of the Degrees of Match

As already mentioned, the matchmaking engines make use of the DoM for ranking the retrieved services. In order to be able to compare these degrees with the corresponding expert relevance assessments we need to express them in a similar form. Specifically, we should define some fuzzy variables that correspond to the various degrees of match. For the reference engine [2] used in this paper, one could define such variables as shown in Figure 6. The name of the linguistic variable *L* related to the degree of match is "DoM" and the set *H(L)* is defined as: *H*("DoM") = {"fail", "subsumed-by", "subsumes", "plugin", "exact"}.

3.3 Generalized Fuzzy Evaluation Measures

One problem, which immediately arises in measuring the effectiveness of a generalized retrieval system, is that a new interpretation should be made for the Boolean "relevant" and "retrieved" sets of service advertisements. Such problem can be resolved through the transformation of the "relevant"/"retrieved" set cardinalities into fuzzy set cardinalities [16]. The proposed measures are generalizations of the recall and precision measures, calculated from the two rankings (i.e., membership functions) of relevance assessments: ***fe*** (delivered by the engine) and ***fr*** (performed by domain experts) in a way similar to the Boolean case[3] (see Eq. 3). Using the fuzzy set cardinalities, the generalized evaluation measures (R_G and P_G), which have been presented in [5], are given in Eq. 4.

$$\boldsymbol{fe}: \mathrm{Q} \times \mathrm{S} \rightarrow [0,1], \quad \boldsymbol{fr}: \mathrm{Q} \times \mathrm{S} \rightarrow [0,1] \qquad \text{Eq. (3)}$$

$$R_G = \frac{\sum_{S_i \in S} \min\{fr(R,S_i), fe(R,S_i)\}}{\sum_{S_i \in S} fr(R,S_i)}$$
$$P_G = \frac{\sum_{S_i \in S} \min\{fr(R,S_i), fe(R,S_i)\}}{\sum_{S_i \in S} fe(R,S_i)} \qquad \text{Eq. (4)}$$

Note that these measures do not represent the proportion of "relevant and retrieved advertisements" to the total number of "relevant" or "retrieved" advertisements, respectively. Since the sets "relevant" and "retrieved" are fuzzy, the evaluation measures take into consideration the membership values of *all* available services and requests. This is indicated by the sums in Eq. 4, that are calculated over all $S_i \in S$.

Another direct observation from the formulae in Eq. 4 is that precision is maximized when the engine estimates are more rigorous than the corresponding expert mappings, i.e., when it holds ***min(fr,fe)=fe*** for the largest proportion of advertised services. The inverse holds for the behavior of R_G.

3.4 Preliminary Experimental Results

In order to compare and outline the differences and potential advantages/disadvantages between EVS1 and EVS2 we have conducted a series of experiments. Some details on them have already been mentioned in Section 2.2 (some statistics are shown in Figure 4). To summarize, we have specified some expert mappings as described in Section 3.1 and have also modeled the engine RSVs as described in Section 3.2. These evaluation results were compared to those of the Boolean scheme. The matchmaking engine we have used is the OWLS-MX Matcher [2], an open source tool. The service collection is the "education" subset of the TC2 collection. The engine was configured to apply only logic-based matching algorithms and the threshold was set to FAIL (i.e., retrieve all logic-based matched services irrespective of their DoM). The precision and recall values for each service request are shown in Table 3.

Table 3. Recall and Precision for EVS1 and EVS2

	EVS1		**EVS2**	
Query ID	$\mathbf{R_B}$	$\mathbf{P_B}$	$\mathbf{R_G}$	$\mathbf{P_G}$
Q15	77%	77%	77%	77%
Q16	60%	92%	87%	96%
Q17	57%	92%	77%	89%
Q18	73%	92%	90%	88%
Q19	100%	65%	100%	71%
Q20	80%	71%	95%	72%

[3] The prefix "f" stands for "fuzzy"

In Q16, there was one service, S_1, "somewhat relevant" but not retrieved and another one, S_2, "irrelevant" but retrieved with the DoM "subsumes". The higher sensitivity of the generalized measures can be demonstrated if we assume that S_2 had been assigned the DoM "exact" and recalculate the precision. Then we observe that while P_B remains unchanged at 92%, P_G decreases from 96% to 93%. Similarly, if S_1 was "very relevant", R_G would decrease from 87% to 84%, while all other measures would remain unchanged.

Another observation from the results of Table 3 is that for the experiments Q16, Q17, Q18, and Q20 there is significant difference between the values of R_G and R_B (with the former being 15% - 27% higher than the latter). In order to explain such behavior we should refer to Figure 4 and notice that these four queries have the highest percentages in the 0.75 deviation in their expert mappings. In other words, the expert has characterized some services as relevant to the query QX, $X \in \{16,17,18,20\}$ in the Boolean case, although a more fine grained mapping would state that the same services are only "slightly relevant" to the query QX.

Finally, in Figure 7 one can see the derived precision-recall plots for the various queries (only three pairs are shown for presentation reasons). The EVS1 behavior is shown through dashed lines and the corresponding EVS2 through solid lines. The other two plots (with 'o' and '+' line styles) depict the average P-R curves calculated based on a macro-evaluation strategy [7]. From these curves it becomes obvious that the EVS2 scheme gives quite analogous results to EVS1 but with higher sensitivity and accuracy, as described previously.

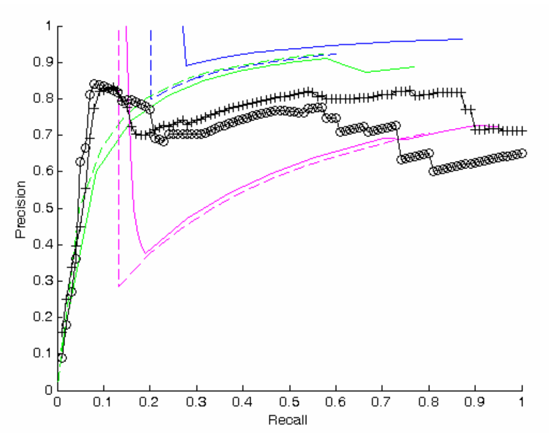


Figure 7. Precision-Recall curves for the experimental dataset (o/dashed line: EVS1, +/solid line: EVS2)

4. A Pragmatic View on Generalized Service Retrieval Evaluation

The generalized evaluation scheme discussed in Section 3 deals with the shortcomings of its Boolean counterpart. However, it raises some serious practical issues. Specifically, most domain experts are not willing to specify so fine-grained mappings between services and requests, since it is a rather demanding task. They would rather prefer to simply give relevance feedback in a Boolean way (i.e., "yes/no"). We believe that every practical evaluation scheme should take this situation into account. Hence, it would be very helpful and interesting if the evaluation scheme could "somehow" infer the expert's mapping value, in a multi-valued system, and properly adjust the given Boolean value. Obviously, such inference is very difficult to achieve, if possible at all. The main reason is that the concept of "relevance" is subjective and depends on many factors. Hence, if one could explain how "relevance" is interpreted by an expert, i.e., identify its components, she would be able to infer some "relevance"-information from a simple Boolean value. Such information could be used for adjusting the Boolean value in a fuzzy way. This hypothesis is depicted in Figure 8, where the value "1" is fuzzified and relaxed to the value "relevant" which is *not* the highest degree of relevance according to Figure 5.

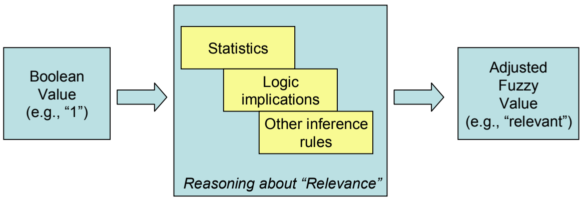

Boolean Value (e.g., "1")
Statistics
Logic implications
Other inference rules
Reasoning about "Relevance"
Adjusted Fuzzy Value (e.g., "relevant")

Figure 8. Automatic adjustment of Boolean expert mapping values

Towards this direction, we assume that a core component of "relevance" is the "logical interpretation of the domain of discourse". That implies that by examining the properties of some "appropriate" logic representation of this domain we could gain more confidence on the relevance between its elements/concepts. However, we remind that logic implication is just one type of inference method. There can also be rules based on user experience, heuristic similarity metrics and any other component that affects the relevance between two concepts, or services in our case. Hence, such inference is a parameterized process.

Furthermore, we assume that a service profile ontology like the one proposed in [3] provides such an "appropriate" logic representation of the services. The concepts of such ontology are complex Description Logic (DL) expressions. For example, if a service S_x takes as input an instance of the class "Book" and returns as output an instance of the class "Price", it could be represented as:

$$S_x \equiv \text{Service} \sqcap \exists \text{ hasInput Book} \sqcap \exists \text{ hasOutput Price}$$

If we perform standard DL classification, which is supported by modern DL reasoners [10], on such ontology we obtain a service profile ontology tree like the one shown in Figure 9. Note that a service request is also a concept in this taxonomy.

Subsequently, we assume that the inference of fuzzy relevance mappings is performed through the inference matrix of Table 4. The first row shows the logic relation between the profile concepts of the request R and a service advertisement S_i. The possible relations are (along with some examples from Figure 9):

Eq: S_i is equivalent to R,
DSup: S_i is direct super-oncept of R (e.g., S_1),
DSub: S_i is direct subconcept of R (e.g., S_6),
Sib: S_i and R are siblings in the service profile ontology (e.g., S_3),
No: no direct relation between S_i and R (e.g., S_x).

The second row shows the existing Boolean expert mappings. Finally, the third row contains the resulting fuzzy "expert" mappings between S_i and R. The notation used in the cells of this row is borrowed from Figure 5. Note that the fuzzy mappings proposed in this matrix have been selected rather intuitively. Future research should aim at optimizing such inference rules. In order to evaluate the effectiveness of the proposed automatic fuzzyfication method we have applied it to the same service collection used in our previous experiment (we call this new scheme EVS2′).

Table 4. Inference matrix

Logic relation	Eq	DSup	DSub	Sib	No
Boolean Value	1	1	1	1	1
Inferred Fuzzy Value	V	R	R	R	SW
Logic DoM	Eq	DSup	DSub	Sib	No
Boolean Value	0	0	0	0	0
Inferred Fuzzy Value	SW	S	S	I	I

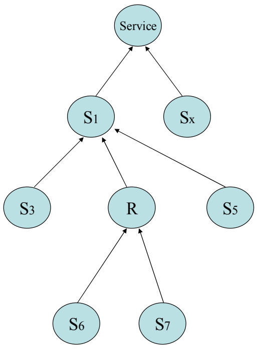


Figure 9. A service profile ontology

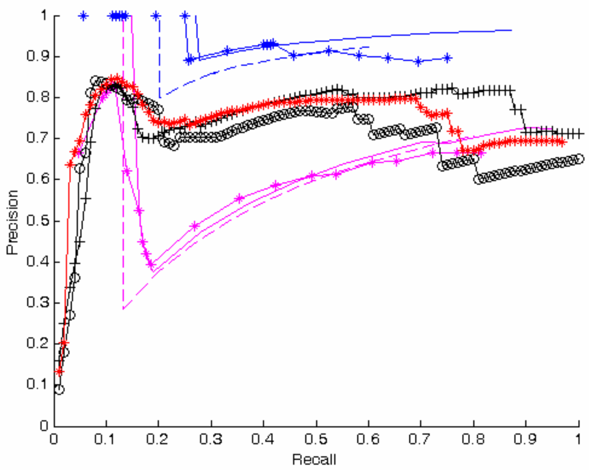


Figure 10. Precision-Recall curves for the automatic fuzzyfication method (o/dashed line: EVS1, +/solid line: EVS2, *: EVS2')

Figure 10 depicts the new average precision-recall curve (with asterisks). This plot approximates the EVS2 curve better that the EVS1 curve. Under the assumption, that the EVS2 evaluation is more accurate than EVS1, we can observe that the proposed inference method is, at least, promising. Hence, we have a preliminary indication that by applying more sophisticated inference algorithms we can compute more "realistic" fuzzy mappings.

Another potential practical problem would be encountered when evaluating a matchmaking engine whose number of degrees of match is different from the number of relevance terms, i.e., when |H("relevance")|≠ |H("DoM")|. In such case, one could use specific fuzzy modifiers (e.g., dilutions, concentrators) over the "relevance" fuzzy values or the "DoM" fuzzy values in order to align them. For instance, consider an engine that assumes H("DoM") = {"exact","plugin","fail"}. Then, the mapping between

the H(“relevance”) and H(“DoM”) sets can be defined as follows:
“fail” is mapped to “irrelevant”,
“exact” is mapped to “very relevant”, and
“plugin” is mapped to *not very* “relevant” *or* (“somewhat relevant” *and not* “slightly relevant”).
The terms *somewhat* and *very* stand for the dilution ($\mu(x) = \mu(x)^{1/2}$) and concentrator ($\mu(x)=\mu(x)^2$) fuzzy modifiers, respectively. In this paper, for reasons of simplicity and due to the used discovery engine, the mapping between the H(“DoM”) and H(“relevance”) sets was one-to-one.

5. Related Work

In this section we briefly review the approaches that have been adopted by other researchers for evaluating the effectiveness of service discovery methods. Note that all of these are Boolean evaluation schemes similar to EVS1, according to the categorization of Table 2. In [8] the standard precision and recall measures are used. Specifically, for each query all the available services are ranked according to some similarity measure (i.e., matching method). The precision and recall values are then calculated for the 50% top-ranked services.

In [11], the authors propose that some variants of precision can capture the precision and ranking quality of the system more precisely. One of them is the Top-k precision (P_k), which is similar to the precision computed in [8]:

$$P_k = \frac{|RT \cap RL|_k}{k}$$

where the nominator denotes the relevant services in the *top k* returned matches. Another measure is R-precision (P_r), which is a variant of P_k where k is substituted by the number of relevant services in the service collection. In addition, the well known recall/precision plot is used, which is widely considered as the most informative graph regarding the effectiveness of matchmaking and search systems.

In [2] the authors exploit the precision-recall curve, too, but calculated through a micro-evaluation averaging technique. Such technique averages the precision and recall values of the various query curves at certain levels λ (for more details the reader is referred to [7]).

6. Conclusion

Semantics in Web Services enable advanced matchmaking and more effective service discovery. In this paper, we discussed on the inadequacy of current evaluation methods to capture the added value provided by such semantics during service discovery. Furthermore, we have proposed a shift from Boolean to generalized evaluation schemes, based on soft computing techniques like Fuzzy Sets. To better justify our thesis, we have conducted some experiments based on an existing service test collection. Finally, we have identified and proposed some possible solutions to some practical issues raised by such evaluation approach. To conclude, although most contemporary SWS research efforts focus on the various aspects of service lifecycle, we believe that more attention should be placed on assessing the real value of the proposed solutions through appropriate, probably new, evaluation schemes. In this paper we have only approached the problem from one possible perspective, namely a fuzzy generalization. Further research should also explore other possible evaluation methods, and especially new metrics.

7. Acknowledgement

This work was performed in the context of the "PENED" Programme, co-funded by the European Union and the Hellenic Ministry of Development, General Secretariat for Research and Technology (research grant 03ED 173).

8. References

[1] Paolucci, M., Kawamura, T., Payne, T. R., and Sycara, K. P., Semantic Matching of Web Services Capabilities*, Lecture Notes In Computer Science: Vol. 2342. First International Semantic Web Conference on the Semantic Web* (pp. 333 – 347), Springer-Verlag, Sardinia, 2002

[2] Klusch, M., Fries, B., Khalid, M., and Sycara, K.. OWLS-MX: Hybrid Semantic Web Service Retrieval. In *1st Intl. AAAI Fall Symposium on Agents and the Semantic Web*, AAAI Press, Arlington VA, 2005

[3] Li, L., and Horrocks, I. “A Software Framework for Matchmaking Based on Semantic Web Technology”, *International Journal of Electronic Commerce*, 6(4), 39-60, 2004

[4] Do, H., Melnik, S., and Rahm, E. “Comparison of schema matching evaluations”, *2nd Annual International Workshop of the Working Group "Web and Databases" of the German Informatics Society (GI)*, Erfurt, Thuringia, Germany, October, 2002.

[5] Buell, D. A. and Kraft, D. H. “Performance measurement in a fuzzy retrieval environment”. 4th Annual international ACM SIGIR Conference on information Storage and Retrieval, Oakland, California, May, 1981.

[6] Bordogna, G., and Pasi, G. “A fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation”, *Journal of the American Society for Information Science*, 44, 70-82, 1993

[7] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. New York: ACM Press, Addison-Wesley, 1999

[8] Stroulia, E., and Wang, Y. “Structural and semantic matching for accessing web service similarity”, *International Journal of Cooperative Information Systems*, 14 (4), 407-437, 2005

[9] OWL-S Service Retrieval Test Collection, http://projects.semwebcentral.org/projects/owls-tc/

[10] Sirin, E., and Parsia, B. “Pellet: An OWL DL Reasoner”. International Workshop on Description Logics (DL2004), Whistler, Canada, 2004.

[11] Dong, X., Halevy, A.Y., Madhavan, J., Nemes, E., and Zhang, J. “Similarity Search for Web Services”. *Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB)* (pp. 372-383), Toronto, 2004.

[12] Zadeh, L. “Fuzzy Sets”, *Information and Control*, 8, 338-353, 1965

[13] Zadeh, L. “The concept of a linguistic variable and its applications to approximate reasoning Part I”, *Information Sciences*, 8, 199-249, 1975

[14] Yager, R. “An approach to ordinal decision making”. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2, 101-113, 1995

[15] Zadeh, L. “From Computing with Numbers to Computing with Words -- From Manipulation of Measurements to Manipulation of Perceptions”, *IEEE Trans. on Circuits and Systems* 45, 1, pp.105-119, Jan. 1999

[16] D. Dubois, A new definition of the fuzzy cardinality of finite sets preserving the classical additivity property, Bull. Stud. Ecxch. Fuzziness Appl. (BUSEFAL) 5, 1981

[17] Tsetsos, V., Anagnostopoulos, C., and Hadjiefthymiades, H., “Semantic Web Service Discovery: Methods, Algorithms and Tools”, chapter to appear in *Semantic Web Services: Theory, Tools and Applications* (Ed. J. Cardoso), Idea Group Publ., March, 2007