Chapter 1
The S3 Contest: Performance Evaluation of Semantic Service Matchmakers

Matthias Klusch

Abstract This chapter provides an overview of the organisation and latest results of the international contest series on semantic service selection (S3). In particular, we introduce its publicly available S3 evaluation framework including the standard OWL-S and SAWSDL service retrieval test collections OWLS-TC and SAWSDL-TC as well as its retrieval performance evaluation tool SME2. Further, we classify and present representative examples of semantic web service matchmakers which participated in the S3 contest from 2007 to 2010. Eventually, we present and discuss selected results of the comparative experimental performance evaluation of all matchmakers that have been contested in the past editions of the S3 series.

1.1 Introduction

In the rapidly growing Internet of services, efficient means for service discovery, that is the process of locating existing services based on the description of their (non-)functional semantics are essential for many applications. Such discovery scenarios typically occur when one is trying to reuse an existing piece of functionality (represented as a web service) in building new or enhanced business processes. Matchmakers [8] are tools that help to connect a service requestor with the ultimate service providers. The process of service selection or matchmaking encompasses (a) the pairwise semantic matching of a given service request with each service that is registered with the matchmaker, and (b) the semantic relevance ranking of these services. In contrast to a service broker, a service matchmaker only returns a rank list of relevant services to the requestor together with sufficient provenance information that allows to directly contact the respective providers. A matchmaker neither composes nor negotiates nor handles the execution of services.

Matthias Klusch
German Research Center for Artificial Intelligence (DFKI), Saarbruecken, Stuhlsatzenhausweg 3, Germany, e-mail: klusch@dfki.de

Semantic matching of services determines the degree of semantic correspondences between the description of a desired service, that is the service request, and the description of a registered service, that is the service offer. For this purpose, both service request and service offer are assumed to be described in the same format. In this chapter, we focus on semantic service matchmakers [7] that are capable of selecting semantic services in formats such as OWL-S[1], SAWSDL[2] or WSML[3], that is services whose functionality is described by use of logic-based semantic annotation concepts which are defined in one or multiple formal ontologies [6]. The processing of such semantic annotations for service selection by a matchmaker bases either on a global ontology it is assumed to shares with service consumers and providers, or on the communication of sufficient ontological information on service annotation concepts to the matchmaker for this purpose. The performance of any service matchmaker can be measured in the same way as information retrieval (IR) systems are evaluated for decades, that is in terms of performance measures like recall, average precision and response time.

Though many implemented semantic service matchmakers exist, there was no joint initiative and framework for the comparative experimental evaluation of their retrieval performance available until few years ago. For this reason, the international contest series on semantic service selection (S3) has been initiated in 2006 by DFKI together with representatives of several other institutions and universities in Europe and USA. Since then it has been organized annually based on a publicly available S3 evaluation framework for semantic service selection which actually consists of the standard test collections OWLS-TC[4] and SAWSDL-TC[5], as well as the evaluation tool SME2[6]. The participation in the contest is by online submission of a matchmaker plugin for the SME2 tool while the final results of each edition of the contest are presented at a distinguished event such as at a major conference of the semantic web or relevant community and/or on the official web site of the S3 contest[7]. The S3 contest series has been exclusively funded by the German ministry of education and research (BMB+F) under project grants 01IW08001 (MODEST, http://www.dfki.de/-klusch/modest/) and 01IW08005 (ISReal, http://www.dfki.de/-klusch/isreal/).

The remainder of this chapter is structured as follows. We briefly introduce the S3 evaluation framework in Section 2. This is followed by a classification of all participants of the contest from 2007 to 2010 together with brief descriptions of some of them as representative examples in Section 3. Eventually, we provide and discuss selected evaluation results in Section 4 before we conclude the chapter in Section 5.

---

[1] http://www.w3.org/Submission/OWL-S/

[2] http://www.w3.org/2002/ws/sawsdl/

[3] http://www.wsmo.org/wsml/wsml-syntax

[4] http://projects.semwebcentral.org/projects/owls-tc/

[5] http://projects.semwebcentral.org/projects/sawsdl-tc/

[6] http://projects.semwebcentral.org/projects/sme2/

[7] http://www.dfki.de/-klusch/s3/

1.2 The S3 Evaluation Framework

The S3 evaluation framework consists of two components: Semantic service retrieval test collections, especially OWLS-TC and SAWSDL-TC, and the S3 evaluation tool for comparative experimental evaluation of retrieval performance of semantic service matchmakers.

1.2.1 Service Retrieval Test Collections

The semantic service retrieval test collections that are currently used in the S3 contests are OWLS-TC and SAWSDL-TC. Both collections are publicly available at the semantic web software portal semwebcentral.org and were initially created at DFKI in 2004 (OWLS-TC1) and 2007 (SAWSDL-TC1). Since their creation these collections have been continuously revised and extended with support of many colleagues from different institutions and universities world wide. Though the collections are commonly considered as standards in the domain of semantic web services today, they do not have the degree of maturity of the prominent TREC (text retrieval conference) collections that have been used in the IR domain for the same purpose for several decades. To achieve this requires a significant increase of joint efforts by the community than it did invest in the building of both collections so far.

OWL-S Service Retrieval Test Collection OWLS-TC. The first version of the collection OWLS-TC [1] was created at DFKI in 2004 and released at semwebcentral.org in 2005. It consisted of only 500 service offers in OWL-S each of which judged by only four users on their binary relevance for only a few service requests in OWL-S. Five years later, the latest edition of the collection, OWLS-TC4, contains already 1083 OWL-S services from multiple domains such as travel, sport, business and healthcare, and 42 queries together with their binary and graded relevance set. These sets of relevant services were determined by more than a dozen users. Each semantic service in OWL-S service of the OWLS-TC4 is grounded in a web service in WSDL 1.1. The semantic annotations of all services base on references to 34 OWL ontologies in total.

Most of the services of OWLS-TC4 were directly retrieved from the web while a few others were created in addition by semi-automated transformation of WSDL services in the web into OWL-S services. Each semantic service in the OWLS-TC is grounded with a WSDL service; for those OWL-S services for which we did not find any grounding in the web, we created one by its semi-automated transformation to WSDL service by use of our OWLS2WSDL tool[8]. In addition, the description of 160 OWL-S service offers and 18 OWL-S service requests in the OWLS-TC4 include logical specifications of preconditions and effects in PDDL 2.0 (Planning Domain and Description Language) and SWRL (Semantic Web Rule Language).

[8] http://projects.semwebcentral.org/projects/owls2wsdl/

The choice of these logical languages was motivated by their widespead use for this purpose and respective recommendations in the OWL-S specification documents. With more than eleven thousand downloads as of February 14, 2011, the OWLS-TC appears to be the by far most widely used semantic service retrieval test collection.

SAWSDL Service Retrieval Test Collection SAWSDL-TC. The retrieval performance of SAWSDL service matchmakers in the S3 contest is measured over the test collection SAWSDL-TC [2]. In its curent version, the SAWSDL-TC3 consists of 1080 services and 42 requests with both binary and graded relevance sets while the semantic annotations of services base on references to 38 ontologies in total. Most of the services of SAWSDL-TC3 were created by semi-automated transformation of all OWL-S services in OWLS-TC3 with the OWLS2WSDL tool (available at semwebcentral.org) to WSDL services which are then transformed to SAWSDL services by manually editing the semantic annotations suggested by the tool without taking service preconditions and effects into account. SAWSDL services that were directly retrieved from the web or have been contributed by the community to this collection currently make up only two percent of its size in total. The collection has been downloaded more than 400 times as of February 14, 2011.

Binary and graded relevance sets of OWLS-TC and SAWSDL-TC. The measurement of the performance of semantic service retrieval requires any test collection to include a set of relevant services for each service request in the collection. Both collections, OWLS-TC and SAWSDL-TC, are providing such relevance sets based on two different types of relevance assessments. The binary relevance of a service offer S to a request R is judged by human users in terms of S being either relevant (relevance score: 1) or not (relevance score: 0). In contrast, the graded relevance of a service is judged by human users on the standard four-graded relevance scale of the NTCIR test collection for IR systems[9]. In this case, the degrees of semantic relevance of a service S to a request R range from ”highly relevant” to ”relevant” and ”partially relevant” to ”not relevant at all” with corresponding relevance scores of 3 to 0. In particular, partially relevant services are assumed to overlap with the service request, that is, the service provides functionality that has been requested and some that has not while the functionality of relevant services is supposed to be subsumed by but not equal to the requested one. The semantic relevance set for each service request is determined by union average pooling of relevance assessments provided by human users: A service S is considered relevant to R if S is judged relevant to R by at least one user, and considered not relevant otherwise, that is, if it is not included in the relevance set of R or has not been rated yet. The graded relevance sets for service requests in OWLS-TC and SAWSDL-TC are available since 2009, respectively, 2010 only.

[9] http://research.nii.ac.jp/ntcir/index-en.html

1.2.2 Evaluation Tool SME2

The evaluation tool SME2 (Semantic Service Matchmaker Evaluation Environment) [4] is part of the S3 evaluation framework. It enables the user to perform an automated comparative evaluation of the retrieval performance of any given set of semantic service matchmakers over given test collections with respect to classical retrieval performance measures.

Retrieval performance measures. The SME2 tool determines the retrieval performance of matchmakers over a given test collection with binary relevance sets by measuring the macro-averaged precision at standard recall levels (MAP), the average precision (AP), the R-precision and precision@k as known from the IR domain. In case of graded relevance, the tool computes the values of the classical measures Q and nDCG (discounted cumulative gain).

For binary relevance: Average precision

$$\mathrm{AP} = \sum_{R \in RS} \frac{1}{|\mathrm{Rel}_R|} \sum_{r=1}^{|L_R|} isrel(r) \frac{count(r)}{r}, \; L_R \text{ rank list of services retrieved for request } R$$

$$isrel(r) = \begin{cases} 1 & \text{if service in } L_R \text{ at rank } r \text{ is relevant} \\ 0 & \text{else} \end{cases}, \quad count(r) = \sum_{i=1}^{r} isrel(i)$$

For binary relevance: Macro-averaged precision at standard recall levels

$$MAP_{\mathrm{n}} = \frac{1}{|RS|} \sum_{R \in RS} \max\{\mathrm{Prec}_{R,o} : \mathrm{Rec}_{R,o} \geq \mathrm{Rec(n)}, (\mathrm{Prec}_{R,o}, \mathrm{Rec}_{R,o}) \in Obs_R\},$$

$$\mathrm{n\text{-}th\ recall\ level\ Rec(n)} = \frac{\mathrm{n}}{\lambda}, \mathrm{n} = 1..\lambda, (\lambda = 20); \mathrm{Ceiling\ interpolation}$$

For graded relevance: nDCG-measure (normalized discounted cumulative gain)

$$nDCG_l = \sum_{R \in RS} \frac{1}{|\mathrm{Rel}_R|} \left(\sum_{r=1}^{\max\{|L_R|, l\}} DCG(r) / \sum_{r=1}^{l} DCG_{ideal}(r) \right)$$

$$DCG(r) = \begin{cases} \dfrac{g(r)}{\log_a(r)} & \text{if } r > a \; (=2) \\ g(r) & \text{else} \end{cases}, \quad \text{Cut-off value } l \; (=100)$$

For graded relevance: Q-measure

$$Q = \sum_{R \in RS} \frac{1}{|\mathrm{Rel}_R|} \sum_{r=1}^{|L_R|} isrel(r) \frac{\omega\, CG(r) + count(r)}{\omega\, CG_{ideal}(r) + r}, \; \omega = 1 \text{ (robustness)}, \; \omega = 0 : Q \cong AP,$$

$$CG(r) = \sum_{i=1}^{r} g(i), \; CG_{ideal}(r) \equiv CG(r) \text{ for perfect service rank list } L_R \text{ for } R$$

Fig. 1.1 SME2: Examples of service retrieval performance measures.

For example, the average precision measure determines for each query how many services retrieved are also relevant and averages this ratio over the set of queries.

The standard recall measure determines how many services which are relevant to a query have been retrieved. In addition, the SME2 tool measures the average query response time, that is the elapsed time per query execution, and the number of accesses by matchmakers to service annotation ontologies, that is their number of respective http-requests during the query-answering test phase. For reasons of space limitation, we omit the comprehensive definition of used performance measures but recall only some of them in figure 1.1 and refer the interested reader to the extensive literature on the topic in the IR domain as well as publicly available software packages on the subject.

SME2 user interface. The SME2 evaluation tool provides an easy to use graphical user interface for configuration, execution and analysis of comparative experimental performance evaluation of any set of pluged in matchmakers over a given test collection. In particular, the tool allows users to tailor summary reports of the evaluation results for archival and printing purposes. Figure 1.2 shows screenshots of the SME2 user interface for an example configuration and the display of evaluation results.

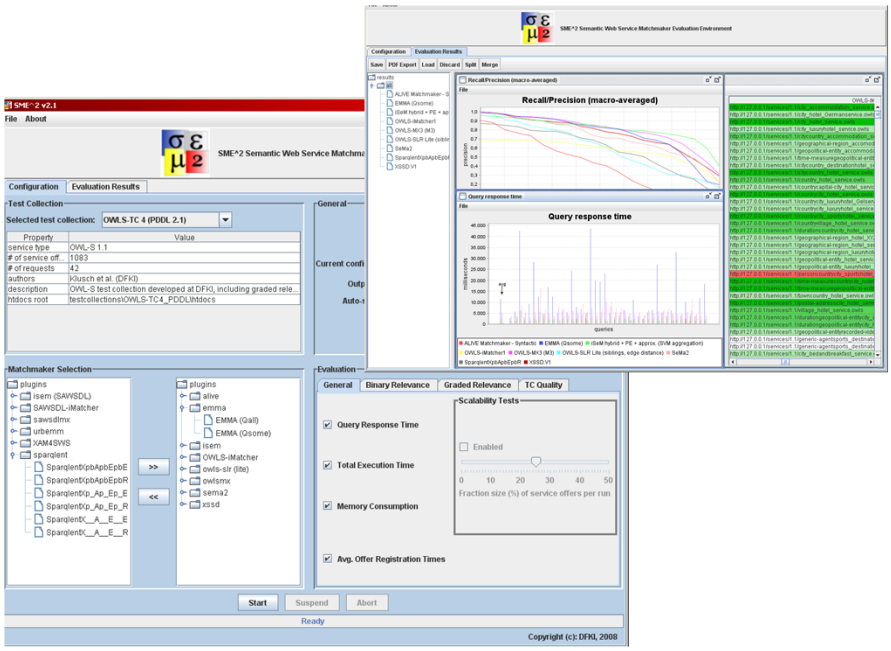

Fig. 1.2 SME2 user interface screenshots for configuration and display of evaluation results.

The SME2 evaluation tool was developed at the German Research Center for Artificial Intelligence (DFKI) and is publicly available since April 2008 at the portal for semantic web software semwebcentral.org. It has been implemented in Java, uses

an embedded Jetty web server and interacts with any semantic service matchmaker through its XML-based plug-in API. The latest version 2.2 of the SME2 evaluation tool also comes with the test collections OWLS-TC4 and SAWSDL-TC3 together with the plugins of all matchmakers that participated in the 2010 edition of the S3 contest.

1.3 Contest Organization and Participants

1.3.1 Organizational Issues

The international S3 contest series was jointly initiated in late 2006 by Matthias Klusch (DFKI, Germany), Ulrich Kuester (University of Jena, Germany), Alain Leger (France Telecom Research, France), David Martin (SRI International, USA), Terry Payne (University of Southampton, UK), Massimo Paolucci (NTT DoCoMo Research Europe, Germany) and Abraham Bernstein (University of Zurich, Switzerland). Since 2007, Patrick Kapahnke and Matthias Klusch at DFKI coordinated four annual editions and presented each of their results at the SMR2 (service matchmaking and resource retrieval in the semantic web) workshops at different locations world wide: 2007 in Busan (South Corea), 2008 in Karlsruhe (Germany), 2009 in Washington D.C. (USA), and 2010 in Shanghai (China).

Actually, the contest consists of two tracks that are devoted to the comparative performance evaluation of matchmakers for OWL-S, and SAWSDL services over the standard service retrieval test collections OWLS-TC and SAWSDL-TC, respectively (cf. Section 2). The contest is open to the comparative evaluation of matchmakers for other kinds of semantic service description as well, if there are respective test collections publicly available for this purpose. This concerns, for example, the evaluation of implemented WSML service matchmakers of which a few are publicly available such as WSMO-MX, but the WSML test collection required for testing them still remains to be created by the community. The same holds for the evaluation of linked (open) services[10] discovery tools like iServe.

The 2009 edition of the contest also offered a special track on an initial cross-evaluation of matchmakers and has been organized by Ulrich Kuester (University of Jena). This track aimed at the evaluation of different matchmakers for different service description formats over respective variants of the same test collection, that is the special collection JGD (Jena Geographic Data Set) of geoservices. We omit the description of the JGD, the evaluation results and lessons learned of this special track since they are presented elsewhere in this book.

The 2010 edition has been organized in collaboration with the semantic technology evaluation campaign of the European research project SEALS project[11] but has not been funded by or made use of any infrastructure of this project. The collabo-

[10] linkedservices.org

[11] http://www.seals-project.eu/

ration included mutual presentations of the initiatives at the international semantic web conference (ISWC 2010).

The participation in the S3 contest is continously open, without any costs and by online submission of a matchmaker code plugin to the SME2 tool together with essential information on the matchmaker itself; the contest web site provides a short guideline of participation. In its first four editions, the S3 contest received 25 submissions in total including 14 OWL-S service matchmakers and 8 SAWSDL service matchmakers. Of these the special track of the 2009 edition received 3 additional submissions, which were IRS-III (Open University, UK), Themis-S (University of Muenster, Germany) and WSColab (University of Modena and Reggio Emilia, Italy). Figure 1.3 summarizes all entries in the order of their first appearance in the contest from 2007 to 2010; all matchmakers except Opossum and ALIVE remained entries for subsequent contest editions.

	Track 1: OWL-S	**Track 2: SAWSDL**	Special Track 2009
2007	OWLS-iMatcher (Kiefer & Bernstein, U Zurich, D) OWLS-MX 1.0 (Klusch & Fries, DFKI, D) JIAC-OWLSM (Masuch, TU Berlin, D)		
2008	OWLS-iMatcher2 (Kiefer & Bernstein, U Zurich, D) OWLS-MX 2.0 (Klusch, Fries & Kapahnke, DFKI, D)	SAWSDL-MX 1.0 (Klusch & Kapahnke, DFKI, D) URBE (Plebani, Politecnico di Milano, I)	
2009	Opossum (Toch, Gal & Dori, Technion, IL) ALIVE (Andreou, U Bath, UK) SPARQLent (Sbodio, HP, I) OWLS-MX3 (Klusch & Kapahnke, DFKI D)	SAWSDL-MX 2.0 (Klusch & Kapahnke, DFKI, D) COM4SWS (Schulte & Lampe, TU Darmstad, D) SAWSDL-iMatcher (Wei & Bernstein, U Zurich, CH)	SAWSDL-MX1/2 SAWSDL-iMatcher IRS-III (Cabral et al, Open U, UK) Themis-S (Müller, U Münster, D) WSColab (Gawinecki, U Modena, I)
2010	SeMa^2 (Masuch, TU Berlin, D) XSSD (Li & Chu, U Beihang, PRC) iSeM (Klusch & Kapahnke, DFKI, D) EMMA (Garcia, Ruiz & Ruiz-Cortez, U Seville, ES) OWLS-SLRlite (Meditskos & Bassiliades, U Thessaloniki, GR)	iSeM-SAWSDL (Klusch & Kapahnke, DFKI, D) LOG4SWS.KOM (Schulte & Lampe, TU Darmstad, D) COV4SWS.KOM (Schulte & Lampe, TU Darmstadt, D)	

Fig. 1.3 Participants of the S3 contest series (showing first appearance only for each).

In the following, we classify all participantsof the contest from 2007 to 2010 according to a classification of semantic service matchmakers proposed in [7].

1.3.2 Classification of Contest Participants

Current semantic service matchmakers including those which participated in the S3 contest can be classified according to (a) the kind of semantic selection they perform, and (b) what parts of the semantic service description they exploit for this purpose. In particular, we may distinguish between means of logic-based, non-logic-based and hybrid semantic service selection based on the service signature (input/output, IO), the specification (preconditions/effects, PE), the full functional profile (IOPE), a monolithic service description in logic, with text or tags, the non-functional service parameters (NF), and combinations of the latter with any part of the functional service description (NF-IOPE). These terms are introduced in Section I of this volume and in [7]. The classification of all contest participants from 2007 to 2010 is summarized in figure 1.4: Since all of them perform semantic matching on either the service signature or the full profile, the figure does not show the remaining levels of the dimension of semantic service description parts. Further, the majority of contested matchmakers performs hybrid semantic matching of service annotations.

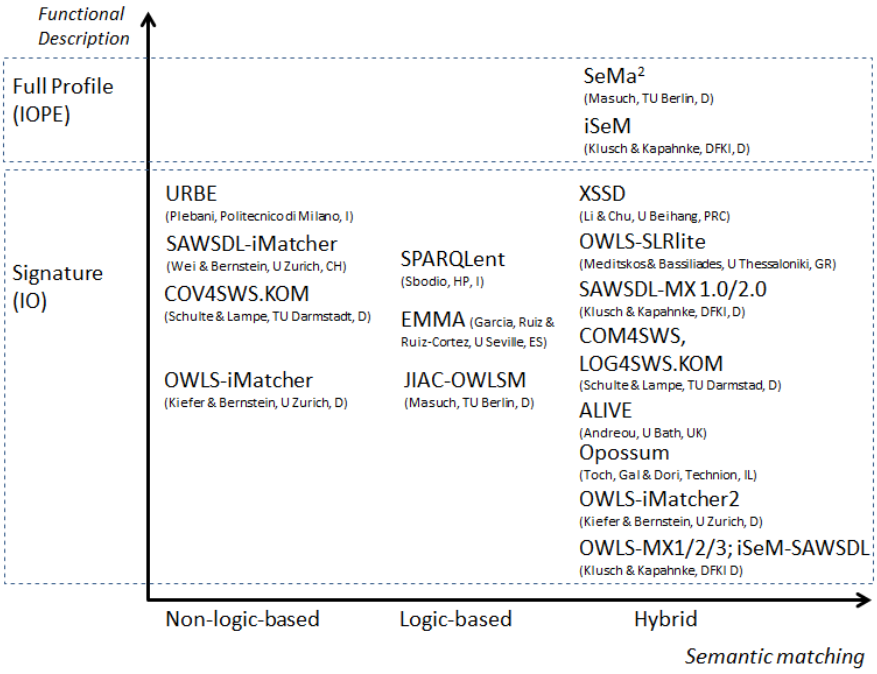

Fig. 1.4 Classification of semantic service matchmakers of the S3 contest from 2007 to 2010.

In the following, we briefly describe selected matchmakers of the contest as representative examples together with an indication of only their best performance results for average precision and response time that were measured in the S3 con-

test. Selected results of the comparative experimental performance evaluation of all matchmakers of the contest will be presented in the next section. For more information on the participants of of the past editions of the contest before 2010, we refer the interested reader to the respective summary reports that are available at the S3 contest web site[12].

1.3.3 Contest Track 1: OWL-S Service Matchmakers

The OWL-S matchmaker track of the S3 contest editions 2007 to 2010 received 14 different OWL-S matchmakers in total. These are ALIVE (University of Bath, UK), OWLS-iMatcher1 and OWLS-iMatcher2 (University of Zurich, Switzerland), Opossum (Technion, Israel), SPARQLent (Hewlet-Packard Research, Italy), XSSD (Beihang University, China), EMMA (University of Seville, Spain), JIAC-OWLSM and SeMa2 (Berlin University of Technology, Germany), OWLS-SLR lite (Aristotle University of Thessaloniki, Greece), and OWLS-MX1, OWLS-MX2, OWLS-MX3 and iSeM (DFKI, Germany).
Most of these matchmakers perform hybrid semantic selection (SeMa2, OWLS-SLR lite, XSSD, Opossum, ALIVE, EMMA) while only two of them focus on non-logic-based, respectively, logic-based semantic selection (OWLS-iMatcher, SPARQLent). Notably, the evaluated OWL-S matchmakers OWLS-MX3, iSeM and OWLS-iMatcher2 perform an adaptive hybrid semantic service selection, that is, they apply machine learning techniques to find the most relevant services to given requests. The matchmaker EMMA does a special kind of logical prefiltering of services before calling the OWLS-MX matchmaker. As representative examples for each of these classes, we now in very brief describe the functionality of of some of these matchmakers together with their best average precision for binary and graded (nDCG measure) relevance and average response times.

Examples of (non-)logic-based matchmakers: OWLS-iMatcher, SPARQLent. The logic-based semantic matchmaker SPARQLent [13] considers the full semantic service profile. It assumes that preconditions and effects are described in SPARQL and computes their query containment relations, in particular RDF entailment rule-based matching of I/O concepts in OWL. SPARQLent has been developed by Marco Luca Sbodio (Hewlett-Packard EIC, Italy) and achieved a high average precision of 0.72 (0.82) with notably low average response time of 0.8 seconds.

The non-logic-based semantic matchmaker OWLS-iMatcher [5] performs signature matching based on the text similarities of service signatures and names. For this purpose, it offers a set of classical token-based and edit-based text similarity measures. OWLS-iMatcher has been developed by Christoph Kiefer and Abraham Bernstein (University of Zurich, Switzerland) and achieved its best average preci-

[12] http://www.dfki.de/ klusch/s3

sion of 0.67 (0.72) in 2.1 seconds of response time in average.

Examples of hybrid matchmakers: XSSD, OWLS-SLRlite, SeMa2, Opossum. The service matchmaker XSSD performs hybrid semantic signature matching based on the computation of logical I/O concept subsumption relations and additional text similarity of service description tags. Service relevance ranking is determined by the degree of logic-based matching followed by the degree of text similarity. XSSD has been developed by Jing Li and Dongjie Chu (U Beihang, China) and achieved a good average precision of 0.79 (0.88) with the second lowest response time of 0.12 seconds measured in the contest (the lowest one was 0.08 seconds measured forthe matchmaker Opossum).

OWLS-SLRlite is similar to XSSD but computes non-logic-based semantic matching scores as aggregated edge and upward co-topic distances between I/O concepts in the respective service annotation ontologies. OWLS-SLR has been developed by Georgios Meditskos and Nick Bassiliades (U Thessaloniki, Greece) and achieved a comparatively low average precision of 0.69 (0.72) but quite fast with an average response time of 0.46 seconds.

In contrast to both matchmakers above, the matchmaker SeMa2 (the successor of JIAC-OWLSM) performs a full profile-based service selection. Its hybrid semantic matching of signatures bases on computing logical I/O concept subsumption relations each of which represented by a fixed numeric score and simple string-based similarity of the concept names. The relevance ranking values are determined through linear weighted aggregation of the scores of both types of matching. Logic-based service sepcification matching by SeMa2 is restricted to the structural comparison of preconditions and effects in SWRL, in particular their instance-based containment relations. Since the service ontologies of the test collections used in the contest do not provide any concept instances (ABox), only the performance of SeMa2's signature matching could be evaluated. SeMa2 has been developed by Nils Masuch (TU Berlin, Germany) and achieved a reasonably high average precision of 0.74 (0.89) but with a high average response time of 4.44 seconds.

The hybrid semantic service matchmaker Opossum combines logic-based with non-logic-based semantic matching of service I/O concepts. The numerical score of the logic-based matching of concepts is combined with the values for their shortest path distance and concept depth (avg. ontology depth) for subsequent ranking. Opossum has been developed by Eran Toch (CMU, USA), Avigdor Gal, Dov Dori (Technion, Israel) and Iris Reinhartz-Berger (Haifa University, Israel). Although its best precision of 0.57 (0.71) is comparatively low it did outrun all other participants in the 2009 edition of the contest with an average response time of 0.08 seconds.
The matchmaker Opossum is described in more detail in a separate chapter of this part of the volume.

Examples of adaptive hybrid matchmakers: OWLS-MX3, iSeM, OWLS-iMatcher2. The matchmakers OWLS-MX3, iSeM and OWLS-iMatcher2 are the only adaptive hybrid semantic matchmakers that have been evaluated in the S3 contest series so far. While OWLS-MX3 [10] restricts its hybrid semantic selection to semantic service signatures, its successor iSeM processes the full (IOPE) profile of

services. OWLS-MX3 adopts the four logic-based I/O concept matching filters of its predecessors OWLS-MX2 and OWLS-MX1 [9], that are logical equal, plugin, subsumes, subsumed-by matching degrees, and applies four different text similarity measures to pairs of unfolded I/O concepts. In addition, it computes numeric scores for ontology-based structural concept similarities. The optimal weighted aggregation of these logic-based and non-logic-based semantic matching filters for relevance decisions and ranking is learned off-line in advance for a given training collections (taken from the test collection) by utilizing a SVM-based (support vector machine) classifier. OWLS-MX3 has been developed by Matthias Klusch and Patrick Kapahnke (DFKI, Germany) and achieved a very high average precision of 0.83 (0.84) but at the cost of a high average response time of 5.37 seconds.

The adaptive hybrid semantic matchmaker iSeM significantly improves upon the OWLS-MX3 by the utilization of additional non-logic-based matching filters and an evidential coherence-based pruning of the given training set during off-line learning of semantic relevance. Notably, iSeM achieved the best average precision of 0.92 (0.84) that has ever been measured in the S3 contest so far, but with only moderately low response time of 2.34 seconds in average. The matchmaker iSeM is described in more detail in a separate chapter of this part of the volume.

The adaptive variant of the matchmaker OWLS-iMatcher, the OWLS-iMatcher2, learns off-line in advance which of its token- and edit-based text similarity measures performs best when applied to pairs of semantic service signatures each of which represented as a weighted keyword vector. OWLS-iMatcher2 also performs a hybrid semantic selection in the sense that it computes the text similarities of signatures based on the logical unfoldings of the respective annotation concepts in the shared ontologies. Notably, OWLS-iMatcher2 achieved the second highest average precision for binary relevance (0.84) in the 2009 edition of the S3 contest.

1.3.4 Contest Track 2: SAWSDL Service Matchmakers

The SAWSDL matchmaker track of the S3 contest editions 2007 to 2010 received 8 different SAWSDL matchmakers in total. These are LOG4SWS.KOM, COV4SWS.KOM and COM4SWS (Darmstadt University of Technology, Germany), SAWSDL-MX1, SAWSDL-MX2 and iSeM-SAWSDL (DFKI, Germany), URBE (Politecnico di Milano, Italy) and SAWSDL-iMatcher3 (University of zurich, Switzerland).
Remarkably, these matchmakers perform either non-logic-based or hybrid semantic service selection. There has been no logic-based semantic service matchmaker submitted to the contest. Since the standard SAWSDL focuses on the annotation of service signatures only, the contested SAWSDL matchmakers perform semantic I/O-based selection only. In the following, we briefly describe some of them as representative examples together with an indication of their best average precision for binary and graded (nDCG measure) relevance and average response times.

Examples of non-logic-based matchmakers: SAWSDL-iMatcher, URBE. The SAWSDL service matchmakers URBE [12] performs non-logic-based signature (I/O) matching based on bipartite graph matching of service (request and offer) operations. In particular, the signature similarity values are computed by means of (a) ontology-based structural I/O concept similarity based on path lengths in a given reference ontology, and (b) WordNet-based text similarity for property class and XSD data type matching of WSDL service elements. The service operation rankings rely on a weighted aggregation of these structural and text matching scores. URBE has been developed by Pierluigi Plebani (Politecnico di Milano, Italy) and has won the SAWSDL track of the contest in 2008 and 2009. It achieved its best average precision with 0.75 (0.85) but also the worst average response time ever measured in this contest (40 seconds).

SAWSDL-iMatcher3 performs non-logic-based signature matching by applying classic vector model-based text similarity metrics to given pairs of semantic annotations of service I/O parameters. This matchmaker has been developed by Dengping Wei and Avi Bernstein (U Zurich, Switzerland), and its best precision of 0.74 (0.85) was as good as URBE but with significantly better response time in average (1.8 seconds).

Examples of hybrid semantic matchmakers: LOG4SWS.KOM, iSeM-SAWSDL. The hybrid semantic matchmaker LOG4SWS.KOM performs first a logic-based matching by computing the logical I/O concept subsumption relations represented as numeric scores. This is complemented by means of non-logic-based matching based on the computation of ontology-based structural I/O concept similarities as the shortest path lengths between concepts and using the WordNet distance as a fallback strategy for missing modelReference-tags of element names. The matchmaker has been developed by Stefan Schulte and Ulrich Lampe (Darmstadt University of Technology, Germany). It achieved the second highest precision of 0.837 for binary relevance and the best precision of 0.89 ever measured for SAWSDL matchmakers in the contest and with even the fastest response time in average (0.24 seconds). The LOG4SWS.KOM matchmaker is described in more detail in a separate chapter of this part of the volume.

The adaptive hybrid semantic matchmaker iSeM-SAWSDL is a variant of the matchmaker iSeM for OWL-S service selection described in the previous section. It processes SAWSDL services and, in contrast to iSeM, performs no specification matching. With an average precision of 0.842 (0.80) this matchmaker won the latest contest by performing slightly better than LOG4SWS.KOM for binary relevance sets, but with an significantly higher response time of 10.7 seconds in average. Notably, despite its adaptation to the given test collection, for the case of graded relevance the matchmaker iSeM-SAWSDL performed worse in terms of precision than all other tested SAWSDL matchmakers. The analysis of this relatively surprising result is ongoing.

1.4 Selected Results of Comparative Evaluation

The comparative experimental evaluation of the retrieval performance of matchmakers that participated in the S3 contest was done by using the SME2 tool and based on the test collections OWLS-TC and SAWSDL-TC (cf. Section 2). In this section, we provide selected results of this evaluation focussing on the potential trade offs between average precision and response time for binary and graded relevance, and those for adaptive and non-adaptive semantic service selection. For more detailed results of all editions and those of the special cross-evaluation track of the 2009 edition of the S3 contest, we refer the interested reader to the respective summary reports available at the S3 contest web site.

Average precision Vs. average response times. The average precisions, nDCG and Q-values, and average query response times of selected OWL-S and SAWSDL matchmakers for editions 2009 and 2010 are summarized in figure 1.5.

2010 (*2009*)		For binary relevance: AP	For graded relevance: nDCG / Q	Response Time AQRT (in secs)
OWL-S matchmakers:				
	iSeM	**0.922**	0.841 / 0.821	2.34
	OWLS-MX3	0.831	0.899 / 0.834	5.37
	SeMa²	0.741	0.83 / 0.73	4.42
Hybrid	XSSD	0.795	0.881 / 0.788	0.12
	ALIVE	0.5	0.42 / 0.64	0.26
	Opossum	0.57	0.71 / 0.51	0.08
	OWLS-SLRlite	0.609	0.723 / 0.57	0.46
Logic	SPARQLent	0.718	0.82 / 0.576	0.57
	EMMA (OWLS-MX2)	0.803	0.881 / 0.815	11.54
Non-logic	OWLS-iMatcher	0.846	0.719 / 0.671	2.15
SAWSDL matchmakers:				
	LOG4SWS.KOM	0.837	0.896 / 0.851	0.24
Hybrid	COV4SWS.KOM	0.823	0.884 / 0.825	0.30
	iSeM-SAWSDL	0.842	0.803 / 0.762	10.66
	SAWSDL-MX1	0.747	0.839 / 0.767	3.86
Non-logic	URBE	0.749	0.85 / 0.777	40.01
	SAWSDL-iMatcher	0.764	0.855 / 0.784	1.79

Fig. 1.5 Best average precisions, nDCG and Q-values with response times.

Please note that the graded relevance sets in the OWLS-TC and SAWSDL-TC are only available since 2009 and 2010, respectively, and some matchmakers like ALIVE and Opossum did not participate in the 2010 edition anymore. We ; we refer to the summary reports of both editions for details. With regard to the measured per-

formance of participants of the S3 contest, there has been a remarkable progress in the development of highly precise semantic service matchmakers since the first edition of the contest in 2007. For example, the best average precision of matchmakers improved from about 0.7 in 2007 to more than 0.9 in 2010. On the other hand, one can observe from figure 1.5 that there still is a large trade off between precision and response time: In both tracks, the fastest matchmakers were not necessarily the most precise ones, and vice versa. Another point to make regarding the results shown in figure 1.5 is that hybrid semantic matchmakers perform generally better than non-hybrid ones in that they offer service selection with a reasonable trade off between precision and response time like XSSD, LOG4SWS.KOM and iSeM.

However, as it has been shown for the popular semantic service matchmaker OWLS-MX in [9], the evaluation results for the COM4SWS matchmaker variants in the 2009 edition of the contest revealed that hybrid semantic matching may also be less precise than mere logic-based matching. This is particularly the case if the latter is complemented by rather than integrated with syntactic matching in terms of, for example, ontology-based structural or text similarity-based semantic filtering of services.

The average query response times of all contested matchmakers largely differed from 40 seconds in the worst case for URBE to 0.08 seconds in the best case for Opossum. Actually, the hybrid matchmaker XSSD whose matching filters are similar to OWLS-MX (cf. Section 3) seems to offer the comparatively best balance of performance by achieving a moderately high degree of average precision (0.79) and running very fast (0.12 seconds). As can be seen from figure 1.5, its precision has been significantly improved only by (a) the adaptive variant of OWLS-MX, that is OWLS-MX3, and (b) the adaptive hybrid matchmaker iSeM which additionally performs ontology-based structural and approximated logical subsumption-based matching - but in any case at the cost of siginificantly higher response times.

The contest results also revealed that the precision of most matchmakers increase when they are measured against the test collections with graded (using both nDCG and Q measures) rather than binary relevance sets. In particular, by using the discounted cumulative gain (nDCG) most matchmakers perform more precisely for the top positions of their service rank lists.

Another lesson learned was that the full-fledged testing of some of the submitted matchmakers was not possible due to the actual characteristics of the test collections. For example, since none of the annotations of services in the OWLS-TC refer to ontologies with assertional knowledge (ABoxes), the matchmakers SeMa2 and SPARQLent could not perform their kind of specification (PE) matching by means of query containment checking over ABoxes. In addition, we observed only a low increase of precision of full-profile (IOPE) matchmakers compared to those which perform signature matching only. The main reason for that is that only 15 percent of the service profiles in the OWLS-TC4 include logical preconditions and effects. In fact, the number of cases in which a matchmaker could avoid false positives of semantic I/O matching by additional checking of service specifications (PE) turned out to be that low such that no significant difference in performance between service IOPE- and IO-based matchmakers has been measured. Thus, the call to the commu-

nity is for extending the test collections with more full service profiles, in particular adding complex logical specifications of preconditions and effects to the semantic annotations of service requests and offers.

With regard to the performance of adaptive and non-adaptive matchmakers, one may observe from the contest results that the former type outperformed the latter in terms of precision in general. Besides, the tested adaptive matchmakers iSeM, OWLS-MX3 and OWLS-iMatcher2 clearly have an edge over the non-adaptive ones by being able to automatically learn offline the best performing combination of their individual matching filters. Though this obviously comes at the cost of prior training time over any given test collection, this apparent disadvantage has to be put into perspective of the potentially huge costs of finding the best filter combination by hand and every time the test collection changes.

Ontology caching strategies. In addition to the selected evaluation results above some further lesson has been learned by mere observation of the tested matchmaker code behavior: Some matchmakers try to reduce their number of time consuming accesses to service annotation ontologies by internal caching of (parts of) ontologies. In fact, according to the results of the experimental http-request behavior analysis of the tested matchmakers with the SME2 tool, one may conclude that different ontology caching strategies did account in part for significant differences in their measured average response times.

For example, the matchmakers XSSD, SeMa2 and OWLS-iMatcher are internally caching the complete set of service annotation ontologies during service registration. This may drastically reduce the number of potentially frequent and high number of external accesses (http-requests) to ontologies during the query processing phase. In fact, the matchmaker only has to get external access to those ontologies it requires to understand the comparatively much fewer service requests of the considered test collection. For the same reason the matchmakers iSeM and OWLS-MX3 are caching not the complete but only the relevant parts of ontologies for matching, that are the logically unfolded, thus self-contained definitions of the service annotation concepts.

1.5 Conclusions

We presented an overview and selected results of the international contest on semantic service selection (S3). The contest runs since 2007 on an annual basis and provides the publicly available S3 evaluation framework that actually includes the standard test collections OWLS-TC and SAWSDL-TC as well as the SME2 tool for automated comparative performance evaluation of semantic service matchmakers.

In general, the progress of development that has been made in the area of semantic service matchmakers is impressive regarding the significant increase of average precision from around 0.6 in 2007 to 0.92 in 2010. Since this increase in precision for most matchmakers is due to the application of more elaborated and complex

matching filters and processing of semantic service annotations, this gain comes at the cost of increased average response time, despite reported streamlinings of matchmaker implementations. The application of ontology caching strategies by some of the contested matchmakers showed promising results in this respect. Actually, the best trade off between precision and speed that has been achieved in the S3 contest is an average precision of around 0.8 with 0.1 seconds response time in average. Hybrid semantic matchmaking appears to be established now, while adaptivity appears to be one of the next trends of development in this domain; it certainly provides the developers with a higher degree of freedom by letting matchmakers automatically learn the best performing combination of given service matching filters.

The S3 contest has been exclusively funded by the German Ministry for Education and Research (BMB+F) in the national projects SCALLOPS, MODEST and ISReal. In 2010 the results of the S3 contest were additionally presented at the first SEALS evaluation campaign workshop. The international S3 contest will be further actively supported on demand by its organizational board (cf. Section 2), in particular by DFKI until 2013 at least.

References

1. OWLS-TC: OWL-S Service Retrieval Test Collection. Latest version OWLS-TC 4.0 (OWLS-TC4) published on September 21, 2010, at semwebcentral: http://projects.semwebcentral.org/projects/owls-tc/. First version of OWLS-TC was created by Benedikt Fries, Mahboob Khalid, Matthias Klusch (DFKI) and published at semwebcentral on April 11, 2005.
2. SAWSDL-TC: SAWSDL Service Retrieval Test Collection. Latest version SAWSDL-TC 3.0 (SAWSDL-TC3) published on September 22, 2009, at semwebcentral: http://projects.semwebcentral.org/projects/sawsdl-tc/. First version of SAWSDL-TC was created by Patrick Kapahnke, Martin Vasileski, Matthias Klusch (DFKI) and published at semwebcentral on July 28, 2008.
3. S. Dietze, N. Benn, J. Domingue, A. Conconi, F. Cattaneo (2009): Two-Fold Semantic Web Service Matchmaking Applying Ontology Mapping for Service Discovery. Proceedings of 4th Asian Semantic Web Conference, Shanghai, China
4. M. Dudev, P. Kapahnke, J. Misutka, M. Vasileski, M. Klusch (2007): SME2: Semantic Service Matchmaker Evaluation Environment. Latest version 2.2 was released on December 2, 2010, at semwebcentral: http://projects.semwebcentral.org/projects/sme2/. First version of SME2 was released at semwebcentral on April 17, 2008. Online support for the SME2 tool is provided by Patrick Kapahnke DFKI.
5. C. Kiefer, A. Bernstein (2008): The Creation and Evaluation of iSPARQL Strategies for Matchmaking. Proceedings of the 5th European Semantic Web Conference (ESWC), Tenerife, Spain.
6. M. Klusch (2008): Semantic Web Service Description. In Schumacher, M.; Helin, H. (Eds.): CASCOM - Intelligent Service Coordination in the Semantic Web. Chapter 3. Birkh”auser Verlag, Springer
7. M. Klusch (2008): Semantic Web Service Coordination. In Schumacher, M.; Helin, H. (Eds.): CASCOM - Intelligent Service Coordination in the Semantic Web. Chapter 4. Birkh”auser Verlag, Springer.
8. M. Klusch, K. Sycara (2001): Brokering and Matchmaking for Coordination of Agent Societies: A Survey. In: Coordination of Internet Agents, A. Omicini et al. (eds.), Chapter 8, Springer.
9. M. Klusch, B. Fries, K. Sycara (2009): OWLS-MX: A Hybrid Semantic Web Service Matchmaker for OWL-S Services. Web Semantics, 7(2), Elsevier.

10. M. Klusch, P. Kapahnke (2009): OWLS-MX3: An Adaptive Hybrid Semantic Service Matchmaker for OWL-S. CEUR Proceedings of 3rd International Workshop on Semantic Matchmaking and Resource Retrieval (SMR2), Washington, USA.
11. M. Klusch, P. Kapahnke, I. Zinnikus (2011): Adaptive Hybrid Semantic Selection of SAWSDL Services with SAWSDL-MX2. Semantic Web and Information Systems, 6(4), IGI Global.
12. P. Plebani, B. Pernici (2010): URBE: Web service Retrieval based on Similarity Evaluation. IEEE Transaction on Knowledge and Data Engineering.
13. M. L. Sbodio, D. Martin, C. Moulin (2010): Discovering Semantic Web services using SPARQL and intelligent agents. Web Semantics, 8(4), Elsevier.
14. E. Toch, A. Gal, I. Reinhartz-Berger, D. Dori: A Semantic Approach to Approximate Service Retrieval ACM Transactions on Internet Technology (TOIT), 8(1), 2007.