Evaluating Semantic Web Service Matchmaking Effectiveness Based on Graded Relevance

Ulrich Küster and Birgitta König-Ries

Institute of Computer Science, Friedrich-Schiller-University Jena
D-07743 Jena, Germany
ukuester|koenig@informatik.uni-jena.de

Abstract. Semantic web services (SWS) promise to take service oriented computing to a new level by allowing to semi-automate time-consuming programming tasks. At the core of SWS are solutions to the problem of SWS matchmaking, i.e., the problem of comparing semantic goal descriptions with semantic offer descriptions to determine services able to fulfill a given request. Approaches to this problem have so far been evaluated based on binary relevance despite the fact that virtually all SWS matchmakers support more fine-grained levels of match. In this paper, a solution to this discrepancy is presented. A graded relevance scale for SWS matchmaking is proposed as are measures to evaluate SWS matchmakers based on such graded relevance scales. The feasibility of the approach is shown by means of a preliminary evaluation of two hybrid OWL-S matchmakers based on the proposed measures.

1 Introduction

In recent years, semantic web services (SWS) research has emerged as an application of the ideas of the semantic web to the service oriented computing paradigm [1]. The grand vision of SWS is to have a huge online library of component services available, which can be discovered and composed dynamically based upon their formal semantic annotations. One of the core problems in the area concerns SWS matchmaking, i.e. the problem of comparing a set of semantic service advertisements with a semantic request description to determine those services that are able to fulfill the given request. A variety of competing approaches to this problem has been proposed [2]. However, the relative strengths and shortcomings of the different approaches are still largely unknown. For the future development of the area it is thus of crucial importance to establish sound and reliable evaluation methodologies. The recent formation of international SWS evaluation campaigns[1] is a promising step in this direction.

One of the core problems of SWS matchmaking is that it is unrealistic to expect advertisements and requests to be either a perfect match or a complete

[1] Semantic Web Service Challenge: http://sws-challenge.org
S3 Contest on Semantic Service Selection:
http://www-ags.dfki.uni-sb.de/~klusch/s3/

fail. Thus, virtually all SWS matchmakers support multiple degrees of match, i.e. they classify the set of advertisements into a hierarchy of different match levels or even assign a continuous degree of match to each offer. Nevertheless, existing approaches for the evaluation of the retrieval effectiveness of matchmaking approaches have so far been based exclusively on binary relevance, i.e. for evaluation purposes an advertisement is considered to be either a match or not, but no further distinction is made. This is a remarkable discrepancy that may distort evaluation results and compromise their reliability. This paper presents an approach to overcome this problem.

The rest of the paper is structured as follows. In the following Section, we provide information about related previous work. In Section 3, we discuss the notion of relevance in the domain of SWS matchmaking and propose a graded relevance scale customized to this domain. In Section 4, we introduce a number of evaluation measures capable to deal with graded relevance. In Section 5, we report on a preliminary experiment on applying the graded relevance scale and the evaluation measures to evaluate two OWL-S matchmakers. We discuss our results with a particular focus on the influence of switching measures and definitions of relevance. Finally, in Section 6, we draw conclusions and outline aspects of future work.

2 Related Work

Experimental evaluation of SWS retrieval has received very little attention so far. The few approaches that were thoroughly evaluated so far exclusively relied on binary relevance and standard measures based on precision and recall. This was also the case with the first edition of the S3 Contest on Semantic Service Selection[2].

The first approach, and the only that we are aware of, to apply graded relevance in SWS retrieval evaluation is the work by Tsetsos et al. [3]. They propose to use a relevance scale based on fuzzy linguistic variables and the application of a fuzzy generalization of recall and precision that evaluates the degree of correspondance between the rating (not ranking) of a service by an expert and a system under evaluation. In this aspect this measure is very similar to the ADM (average distance measure) measure proposed by [4]. Unlike measures that evaluate the ranking created by a retrieval system these measures evaluate the absolute score assigned to a retrieved item by the system. This can lead to counterintuitive results since such measures are obviously biased against systems that rank services correctly but generally assign relatively higher or lower scores [5]. The measures that we use in this work avoid this issue.

Di Noia et al. obtained reference rankings for service matchmaking evaluations by directly asking human assessors to rank the available services [6]. This approach avoids the imprecision related to binary relevance judgments and generally yields more stable results than inducing a reference ranking via relevance judgments. However, it also requires much more effort from the human

[2] http://www-ags.dfki.uni-sb.de/~klusch/s3/

assessors and is thus difficult to scale to large datasets. Di Noia et al. evaluate the matchmaking performance using rank correlation measures from statistics. These measures estimate the difference between two rankings but, for instance, do not differentiate whether the rankings differ in the top ranks or the bottom ranks. Yet, for most retrieval settings, the correctness of the top ranks is much more important than that of the bottom ranks. The measures proposed in this work allow to take such considerations into account.

There is a large body of related work from the area of Information Retrieval that concerns the development of measures based on graded relevance as well as investigations of their properties [5, 7–12]. We rely heavily on these achievements and our work can be viewed as an application and adaptation of this work to the SWS retrieval evaluation domain. We are not aware of any previous work on relevance schemes specifically designed for the SWS retrieval domain and discussions on how to provide reliable and consistent relevance judgments within this domain.

3 Relevance for SWS Retrieval Evaluation

The criteria most often used for experimental retrieval evaluation has been the effectiveness of a retrieval system, i.e. how good a system is in retrieving those and only those items that a user is interested in. Effectiveness evaluations are thus based on the notion of *relevance* of an item to a query [13]. Most evaluation campaigns, in particular TREC[3], have primarily been based on binary relevance, i.e. a document (in the terminology of TREC) was considered to be either relevant or irrelevant to a topic, but no further distinction was made.

The few attempts for quantitative SWS retrieval effectiveness have so far adopted this binary approach [14–17]. However, it has been argued that binary relevance is too coarse-grained to evaluate SWS matchmaking approaches [3, 18]. This view is supported by the fact that nearly all SWS matchmaking algorithms are designed to support multiple degrees of match (DOMs). In a classic paper, Paolucci et al., for instance, proposed the use of *exact*, *plug in*, *subsumes*, and *fail* [19]. This scale or variations thereof have been adopted by many approaches.

It is thus desireable to employ a graded relevance scale instead of a binary one in SWS retrieval evaluations. However, the design of such a scale is far from trivial.

To be practically useful it must have clear definitions that enable domain experts to provide reference relevance judgments as unambiguously as possible. In this aspect a scale like *very relevant*, *relevant*, *somewhat relevant*, *slightly relevant*, and *irrelevant* as used by [3] is very difficult to judge objectively. On the other hand, human assessors should judge the relevance of a service offer with respect to a service request on the level of the original services and not their semantic formalizations. After all, the appropriateness and quality of these formalizations is also part of what is being evaluated. It is therefore not appropriate to directly use the DOMs by Paolucci et al. as a relevance scale for

[3] http://trec.nist.gov/

general SWS retrieval evaluation either. The definition of these DOMs is only meaningful in the context of DL subsumption reasoning, i.e. in the context of a particular formalization approach. It can not be meaningfully applied outside of this context.

To define a relevance scale that is equally applicable to different approaches but still sufficiently well defined to allow objective judgments, some assumptions and central terms need to be clarified. To this end, we recapitulate the basic definitions from a conceptual architecture for semantic web services presented by Preist [20]. According to this architecture, a service is defined as a *provision of value* in some given domain, e.g. the booking of a particular flight ticket. Web services are technical means to access or provide such services. Service providers typically do not offer one particular service, but a *set* of coherent and logically related services, e.g. booking of flight tickets in general and not a specific flight ticket. Service descriptions will thus describe the set of services a provider is able to offer respectively a requester is interested in. Due to dynamics involved, privacy issues, and limited precision and detailedness, service descriptions will not always precisely capture the set of services that a provider is able to deliver or that a requester is interested in. Instead, they may be incorrect (not all described services can be provided or are of interest) as well as incomplete (not all services that can be provided or are of interest are covered by the description).

Keller et al. extended this model by remarking, that descriptions based on this model are not semantically unambiguous without knowing the intention of a modeler, which can be that either all or only some of the elements contained in the described service set are requested respectively can be delivered [21]. Based upon this consideration they formally define different set theoretic match relationships between service offer and request descriptions. Because of its flexibility combined with clear definitions and its grounding to a well-defined conceptual model we propose a relevance scale that builds upon the match relationships introduced by Keller et al., extended by the notions of RelationMatch and ExcessMatch that we will explain below:

Match: The offer satisfies the request completely.

PossMatch: The offer might satisfy the request completely, but due to the incompleteness of the descriptions this can not be guaranteed based upon the available information.

ParMatch: The offer satisfies the request, but only partially (it offers some of the services which are requested but not all).

PossParMatch: The offer might satisfy the request partially, but due to the incompleteness of the descriptions this cannot be guaranteed based upon the available information.

RelationMatch: The offer does not provide services as requested, but related functionality. Thus, it could be useful in coordination with other offers.

ExcessMatch: The offer is able to provide the requested services but would result in additional undesirable effects that are not requested by the client.

NoMatch: None of the above, the offer is completely irrelevant to the request.

As a first remark, please note that these relevance degrees are not totally ordered. It will depend on the particular use case at hand, whether e.g. a definite partial match is preferable or not to a possible full match. Match, PossMatch, ParMatch, PossParMatch, and NoMatch have been introduced by Keller et al. We omit a detailed discussion due to space limitations and refer the interested reader to [21]. Instead, we will focus on RelationMatch and ExcessMatch and motivate why these extensions are necessary.

A ParMatch characterizes a situation where the client requests multiple services and a provider is capable of delivering only some of those. A similar situation arises, when, for instance, a web service is able to deliver the desired effects, but the client is unable to provide the required inputs. Consider for instance a web service able to provide flight bookings between airports identified by the international airport code and a client that requests a flight between two particular cities. The web service can not be used directly to fulfill the client's request but intuitively it would still constitute a partial match. Such situations may arise in the context of all of the four typical elements of services: inputs, outputs, preconditions and effects. To distinguish such advertisements from completely irrelevant ones, but also from the clear defined ParMatch, we added the notion of RelationMatch.

We continue with a discussion of ExcessMatch. Typically, a full match between a service advertisement and request is defined as meeting the following conditions [2]: All inputs required by the offer are available, the preconditions of the advertisement are satisfied by the state of the world prior to the service execution, and the offer provides all outputs and effects required by the client. The first two conditions concern the applicability of a service in a given situation, the last concerns its usefulness with respect to the client's request. Most approaches disregard a problem that arises, if a web service delivers more effects than are requested by the client. A client wanting to purchase a cell phone (only requested effect) would likely reject an advertisement that sells a cell phone (Effect 1) bundled with a contract with a specific telecommunication company (Effect 2). Nevertheless, most SWS matchmaking approaches would consider this a perfect match since all requested effects are delivered by the provider at hand. Similarly, a client looking for apartments in Berlin may or may not accept a web service providing a listing of apartment offers if that listing can not be restricted to offers located in Berlin. To accommodate such situations, we added the notion of ExcessMatch.

Finally, we would like to point out that, strictly spoken, the differentiations between Match and PossMatch (level of guarantee in the presence of imprecise descriptions), ParMatch and Match (level of horizontal completeness), RelationMatch and Match (issue of partial incompatibilities), and ExcessMatch and Match (issue of unwanted additional effects) are actually four unrelated dimensions that would result in 16 (2^4) levels of relevance even if each dimension is considered to be binary. To keep relevance levels manageable by the domain experts providing reference judgments, we restrict the scale to the seven relevance levels listed above for the time being. These seem to be the most important, but

a further investigation of the optimal number of relevance levels is necessary and will be done in future work.

4 Evaluation Measures Based on Graded Relevance

To leverage the extra information contained in graded relevance judgments and graded degrees of match in a retrieval effectiveness evaluation, the retrieval measures for binary relevance need to be generalized to graded relevance. In this section, we present such generalized measures. To make the paper self-contained, we start by briefly recalling some basic definitions for the binary case.

Throughout this paper, we use the following definitions. Let R be the set of relevant items for a query. Let L be the set of items returned in response to that query. Then *Recall* is defined as the proportion of relevant items returned ($Recall = \frac{L \bigcap R}{R}$) and *Precision* as the proportion of returned items that are relevant ($Precision = \frac{L \bigcap R}{L}$).

Recall and Precision are set-based measures. However, there is an obvious trade-off between them. By returning more items, a system can usually increase its Recall at the expense of its Precision. Thus, in the following we assume that systems return a ranked output ordered by estimated confidence in relevance. Let $r, 1 <= r <= |L|$ denote a specific rank in this output. Let $isrel(r) = 1$, if the item at rank r is relevant and 0 otherwise. Let $count(r)$ be the number of relevant items among the top r retrieved items, i.e. $count(r) = \sum_{i=1}^{r} isrel(i)$.

This allows to measure Precision as a function of Recall by scanning L from the top to the bottom and measure the Precision at standard Recall levels. These measures average well for different queries and the corresponding R/P charts are the most widely used measure to compare the retrieval performance of systems. It is also possible to measure Precision and Recall at predefined ranks ($Precision_r$ and $Recall_r$, r is often referred to as document cutoff level). However, these measures do not average well for queries where $|R|$ varies greatly.

If a system's performance needs to be captured in a single measure, the probably most often used one is *Average Precision* over relevant items which is defined as: $AveP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{count(r)}{r}$.

Since about 2000, there is an increased interest in measures based on graded or continous relevance. Various proposals have been made to generalize the measures introduced above from binary to graded relevance (see [12] for a discussion). Most of these are based on or can be expressed in terms of *Cumulated Gain* proposed by Järvelin and Kekäläinen [8]. Intuitively, Cumulated Gain at rank r measures the gain that a user receives by scanning the top r items in a ranked output list. More formally let $g(r) >= 0$ denote the gain value (or the relevance level) of the item at rank r and from now on $isrel(r) = 1$, if $g(r) > 0$ and 0 otherwise. Then Cumulated Gain at rank r is defined as $cg(r) = \sum_{i=0}^{r} g(r)$. Moreover consider an ideal ranking, i.e. $\forall(r > 1, r <= |R|) : isrel(r) = 1$ and $\forall(r > 1) : g(r) <= g(r-1)$. Let $icg(r)$ (*Ideal Cumulated Gain* at rank r) denote the Cumulated Gain for this ideal ranking.

Since $cg(r)$ can take arbitrarily large values for queries with many relevant items it has to be normalized to average or compare results across queries. *Normalized Cumulated Gain*[4] at rank r is defined as the retrieval performance relative to the optimal retrieval behavior, i.e. $ncg(r) = \frac{cg(r)}{icg(r)}$.

It allows a straightforward extension of AveP which has sometimes been referred to as Average Weighted Precision [5]: $AWP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{cg(r)}{icg(r)}$.

Unfortunately, $ncg(r)$ has a significant flaw that AWP inherits: since $icg(r)$ has a fixed upper bound ($icg(r) <= icg(|R|)$), $ncg(r)$ and AWP cannot penalize late retrieval of relevant documents properly since $ncg(r)$ cannot distinguish at which rank relevant documents are retrieved for ranks greater than R [11]. This can be illustrated by comparing $ncg(r)$ and $Precision_r$ for the last rank in a full output ($R \subseteq L$). In this case $ncg(|L|) = 1$ but $Precision(|L|) = \frac{|R|}{|L|}$, which is usually much smaller than one. Several measures have been proposed that resolve this flaw of AWP.

Järvelin and Kekäläinen [8] suggested to use a discount factor to penalize late retrieval and thus reward systems that retrieve highly relevant items early. They defined *Discounted Cumulated Gain* at rank r as $dcg(r) = \sum_{i=0}^{r} \frac{g(r)}{disc(r)}$ with $disc(r) >= 1$ being an appropriate discount function. Järvelin and Kekäläinen suggest to use the log function and use its base b to customize the discount which leads to $disc(r) = \log_b r$ for $r > b$ and $disc(r) = 1$ otherwise (the distinction is necessary to maintain $disc(r) >= 1$ to avoid boosting the first ranks).

We use an according definition of *Ideal Discounted Cumulated Gain (idcg(r))* to define an adapted Version of AWP that we call *Average Weighted Discounted Precision*:

$$AWDP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{dcg(r)}{idcg(r)}.$$

Similarly, Kishida [12] proposed a generalization of AveP that also avoids the flaw of AWP:

$$genAveP = \frac{\sum_{r=1}^{|L|} isrel(r) \frac{cg(r)}{r}}{\sum_{r=1}^{|R|} \frac{icg(r)}{r}}$$

Furthermore, Sakai [5] proposed an integration of AWP and AveP called Q-measure which inherits properties of both measures and possesses a parameter β to control whether Q-measure behaves more like AWP or more like AveP:

$$Q\text{-}measure = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{\beta cg(r) + count(r)}{\beta icg(r) + r}$$

All measures allow to finetune the extent to which highly relevant items are preferred over less relevant items (by setting appropriate gain values) but differ in the degree of control that is possible with respect to the extent to which

[4] A similar measure has been proposed by Pollack in 1968 under the name *sliding ratio*.

late retrieval is penalized. Q-Measure controls the penalty by its β Parameter, $AWDP$ by the choice of an appropriate discounting function, and genAveP lacks such control. Sakai [9] discusses this issue in detail but unfortunately disregards choices of disounting functions for $ndcg(r)$ besides the logarithm.

5 Experimental Retrieval Evaluation

We now report on the evaluation of our appraoch by means of a preliminary experiment on using the relevance scale introduced in Section 3 and the measures introduced in the previous section to evaluate the retrieval effectiveness of two matchmakers. We start by describing the test data we used and in particular our experiences on obtaining graded relevance judgments. We continue by describing the parameters that we chose for the experiment and complete our report with a discussion of our results.

5.1 Test Data

Unfortunately, there is still a lack of *standard* test collections in the area of SWS [18]. To test the proposed evaluation approach, we chose the Education subset of the OWLS-TC 2.2 test collection[5]. This subset contains 276 OWL-S service descriptions and six request descriptions together with binary relevance judgments. We chose this subset mainly for two reasons. First, this subset[6] had been used previously in an experiment with graded relevance judgments which allows to compare our results with the results from that previous experiment [3]. Second, for OWLS-TC, ranked outputs from two different matchmakers, OWLS-M3 [14] and iMatcher [16], are available through the organizers of the S3 Matchmaker Contest[7]. However, it turned out that iMatcher was unable to process one of the six queries which was thus excluded from the test data. Further information including all test data and results are available online[8].

To collect and manage graded relevance judgments for this subset, we used the OPOSSum portal[9] which already lists all the OWLS-TC services. Therefore, throughout this paper we identify queries by their id from that portal (5654, 5659, 5664, 5668, and 5675). We extended OPOSSum with a user interface that allows to conveniently enter graded relevance judgments for large numbers of services. We developed some guidelines for relevance judges[10] and three persons (one expert in the area of SWS as well as two volunteers that had only a basic understanding of SWS) judged the complete subset.

Unfortunately, it turned out that the judgments of the three judges did not correspond with each other very well. We believe that this is largely caused by

[5] http://projects.semwebcentral.org/projects/owls-tc/
[6] More precisely a similar subset from a smaller previous release of this test collection.
[7] http://www-ags.dfki.uni-sb.de/~klusch/s3/
[8] http://fusion.cs.uni-jena.de/OPOSSum/ISWC08-SMRR/
[9] http://fusion.cs.uni-jena.de/OPOSSum
[10] http://fusion.cs.uni-jena.de/OPOSSum/index.php?action=relevanceguideline

	Match	Poss	Par	PossPar	Relation	Excess	None
Relevant	130	12	33	5	6	-	20
Irrelevant	8	3	7	1	37	-	1408
Average	0.94	0.8	0.83	0.83	0.14	-	0.01

Table 1. Correspondance with binary OWLS-TC 2.2 judgments

	Match	Poss	Par	PossPar	Relation	Excess	None
Very r.	24	1	4	0	0	-	0
Relevant	19	1	2	0	0	-	2
Slightly r.	11	7	1	0	1	-	1
Somewhat r.	10	2	3	2	2	-	4
Irrelevant	3	0	0	1	0	-	15
Average	2.75	1.64	2.9	1.33	1.5	-	0.68

Table 2. Correspondance with fuzzy judgments by Tsetsos et al.

insufficient textual documentation of the services in the employed test collection. This lack of detail required relevance judges to make a lot of assumptions regarding the semantics of the services. Consequently, single judges were able to judge consistently but judgments varied between the judges depending on the different assumptions that were made (for instance whether a lecturer or a research assistant are considered researchers or not). For the rest of this paper and the reported preliminary experiment we used the judgments of the SWS expert exclusively.

We compared these judgments with the binary OWLS-TC judgments. Table 1 shows that correspondance. For each graded relevance level it shows how many of the services judged into this level were evaluated relevant versus irrelevant by the OWLS-TC authors. The average row shows the arithmetic mean that is computed by assigning a value of one/zero to the binary relevant/irrelevant services. Please note that none of the services in the Education subset of OWLS-TC was judged an ExcessMatch by our judges. Nevertheless we believe that this relevance level has its own right of existence for other collections.

Since OWLS-TC employs a very liberal definition of relevance, we were surprised to see eight services judged irrelevant by OWLS-TC but judged a perfect Match by our judges. A closer look revealed that seven of those eight mismatches seem to indicate errors in the OWLS-TC reference judgments. The remaining mismatch is caused by different context knowledge assumptions. Such assumptions also explain most of the other mismatches, like the twenty services judged irrelevant by us but relevant by OWLS-TC. Most of these, for instance, are related to a query for scholarships. Services providing information about loans were judged relevant by OWLS-TC but irrelevant by our expert.

Finally, we compared our judgments with the fuzzy relevance judgments made by Tsetsos and colleagues [3] for the OWLS-TC 2.1 Education subset, which contains the same requests as the 2.2 subset but only 135 instead of 276 services. Tsetsos et al. used a fuzzy scale with the values irrelevant, slightly rele-

vant, somewhat relevant, relevant, and very relevant. For each graded relevance level Table 2 shows how many of the services judged into this level by our judges were judged into each of their fuzzy levels by Tsetsos et al. The average values are computed by assigning values of zero through four to the relevance levels used by Tsetsos et al. The small numbers in the Irrelevant row are caused by the fact that we used only explicit judgments, but Tsetsos et al. provided most "irrelevant" judgments only implicitly. Thus, with a full set of explicit judgments, numbers in the Irrelevant row would have been much higher and the Averages in particular in the last column much lower. We were surprised that the services judged as a perfect Match by our judges were relatively evenly distributed among the four top relevance levels of Tsetsos et al. (see first column). Since we could not obtain information about the rationale behind those judgments or the precise definitions of the relevance levels we lack an explanation for this phenomena but we expect it to be caused by the same issues that caused our judges to judge differently relatively often, too.

5.2 Evaluation Parameters

The measures described in Section 4 allow to evaluate SWS retrieval systems based on the graded relevance scheme introduced in Section 3 but leave open the question about the proper parameter combinations to use in an evaluation. As Järvelin and Kekäläinen remark, "the mathematics work for whatever parameter combinations and cannot advise us on which to choose. Such advise must come from the evaluation context in the form of realistic evaluation scenarios" [8]. In order to perform an investigation in particular of the effects of switching from binary to graded relevance, we chose two gain value settings that actually correspond to binary relevance and two settings which leverage the potential of graded relevance. The corresponding gain values are displayed in Table 3. *Strict Binary* and *Relaxed Binary* correspond to strict versus relaxed definitions of binary relevance. *Graded 1* corresponds to a setting with a focus on high precision which is appropriate in a use case of automated dynamic binding whereas *Graded 2* reflects a more balanced preference between precision and recall which seems more appropriate in use cases where a human programmer is searching for a service. Additionally (not shown in Table 3) we used the binary relevance judgments that come together with OWLS-TC 2.2 for comparison.

For each of the five queries and each of the five gain value settings, we computed the following measures for both matchmakers: AWDP using the discount functions r (AWDP-R), $\sqrt{r}$ (AWDPSQRT), $\log_2 r$ (AWDPLog2) and $\log_{10} r$ (AWDPLog10) as well as without discount function (AWP), Q-Measure with $\beta = 5$, $\beta = 1$, $\beta = 0.5$, and $\beta = 0$, and genAveP. Using a quickly growing discount function in conjunction with AWDP (e.g. AWDP-R) rewards systems that retrieve highly relevant items early, i.e. it puts the emphasis of the evaluation on the top ranks. Using no discount function (AWP) leads to a more balanced consideration of all ranks at the prize of loosing the ability to penalize a very late retrieval of items. Slowly growing discount functions (e.g. AWDPLog2) constitute a compromise between these extremes. In the case of Q-Measure a larger

	Strict Binary	Relaxed Binary	Graded 1	Graded 2
Match	1	1	6	4
PossMatch	0	1	2	2
ParMatch	0	1	1	2
PossParMatch	0	1	0.5	1
RelationMatch	0	1	0	2
ExcessMatch	0	1	0	1
NoMatch	0	0	0	0

Table 3. Experimental gain value settings

β makes Q-Measure more similar to AWP, i.e. rewards retrieving highly relevant items prior to marginally relevant items but makes it vulnerable to not penalizing very late retrieval of relevant items. Small choices for β make Q-Measure more similar to the traditional binary AveP which does not differentiate between highly and marginally relevant items but correctly penalizes late retrieval of relevant items. A choice of $\beta = 0$ completely reduces Q-Measure to binary AveP. Similarly, genAveP is reduced to AveP in settings with binary relevance.

5.3 Results

As expected, results vary significantly over queries. For Query 5675, for instance, M3 is rated higher by 40 out of the 50 possible combinations of measures and gain value settings. In contrast, for Query 5675 iMatcher is rated higher by all measures. Given this large variation, the relatively small size of our data set and in particular the fact that we had data only from two matchmakers, the results that we report in the remainder of this section need to be taken with a grain of salt. Nevertheless, they indicate some interesting preliminary findings.

Our results confirm the expectation, that the choice of measure matters, not only in terms of absolute numbers but also in terms of which matchmaker is rated higher. This is illustrated by Figure 1 that shows the values of the various measures for Request 5654 and Strict Binary and Graded 1 gain value settings. For this request, AWDP with large discounting favors iMatcher while AWDP with little or no discounting as well as Q-measure favor M3.[11]

While results frequently changed with different measures, we found that, except for $\beta = 0$, the choice of β has little influence on the absolute and relative performance of the matchmakers (see Figure 1). In fact, with our data, different parameterizations for Q-measure almost never made a difference in terms of which matcher is rated higher. Furthermore, genAveP always ranked the two matchmakers the same way Q-measure did.

For the binary cases, this behavior of Q-measure can be well explained. In this case $cg(r) = count(r)$ and $icg(r) = r$ if $r <= |R|$. Thus, Q-measures fraction can

[11] Please note that this finding (Q-measure favors M3) are specific to this request. We found frequent changes of the favored matchmaker when changing the measure but no general trend that a particular measure favors a particular matchmaker.

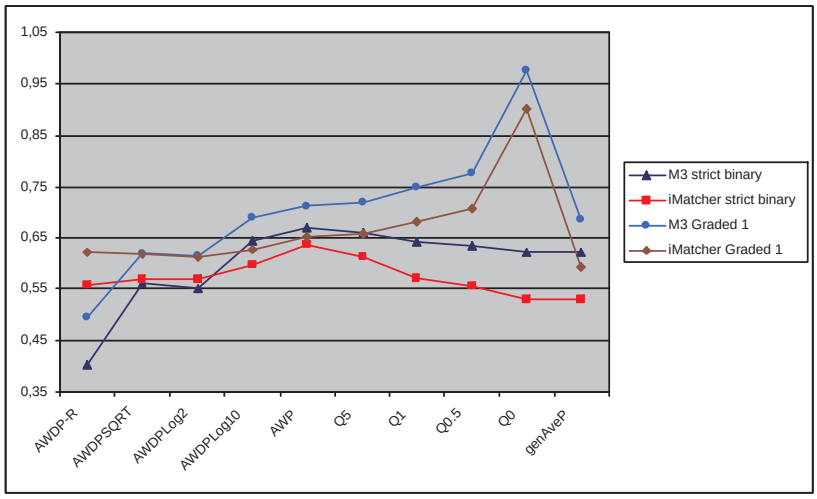

Fig. 1. Results for Request 5654 with Strict Binary and Graded 1 gain values.

be reduced by $\beta+1$ if $r <= |R|$. In other words, in binary cases, β influences the value of Q-measure only for relevant items retrieved after rank $|R|$. Relatively few relevant items are retrieved at such ranks in our experiment, thus the influence of β to the value of Q-measure is limited with our data.

Compared to the influence of Q-measure's β, the choice of discount function of AWDP caused more changes in the ratings. It didn't cause changes in the ratings for Queries 5664, 5668, and 5675 but for the remaining two queries the different versions of AWDP disagreed in eight of ten cases (two queries times five gain value settings), including those displayed in Figure 1.

The notable peak of both matchmaker's performance for the Graded 1 gain value setting measured with Q0 that is visible in Figure 1 highlights how the use of graded relevance influences measure results. Using $\beta = 0$ reduces Q-measure to AveP and thus Graded 1 to a binary scale which largely resembles the original OWLS-TC judgments. For both matchmakers, this results in a significantly increased absolute performance, albeit not in a change of their performance relative to each other.

Generally, changes in the settings of the gain values caused more significant changes in how the matchmakers were rated than changes in the parameterizations of AWDP and Q-measure. However, the Q-measure variations and genAveP were again less sensitive towards changes in the evaluation parameters than the AWDP-family. Their ratings did not change regardless of the gain values used except for Query 5664 where they all preferred M3 with the Strict Binary and iMatcher with the other settings[12]. In contrast, with the one exception of Query

[12] Except for Q-measure with $\beta = 5$ and Graded 2. This case favored M3, too.

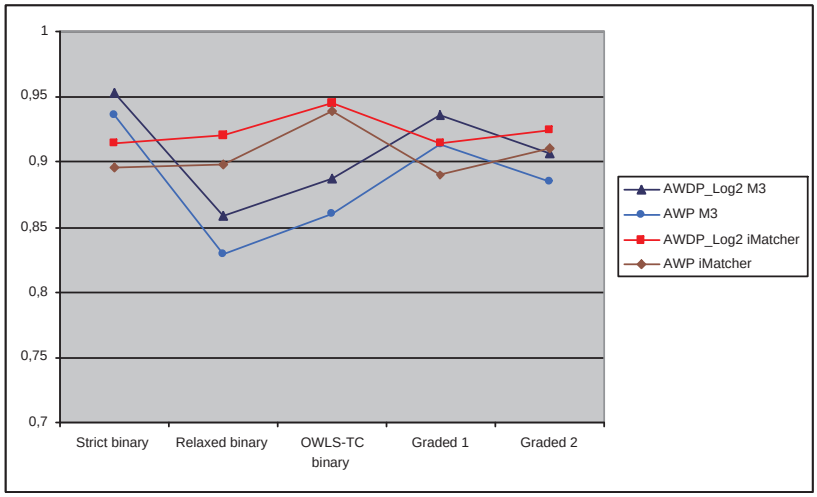

Fig. 2. AWDP measures for Request 5664 for different gain values.

5668 where iMatcher outperforms M3 regardless of the measure, changes in the gain value settings caused changes in the ratings of the AWDP measure family in nearly half of the cases. As an example, Figure 2 shows the values of AWDPLog2 and AWP for Request 5664: M3 is favored by both measures for Strict Binary and Graded 1 while iMeasure is favored for Relaxed Binary, OWLS-TC Binary, and Graded 2. Generally, Relaxed Binary and OWLS-TC Binary tend to benefit iMatcher while the other settings tend to benefit M3. The most likely interpretation is that M3 performs a stricter selection and thus outperforms iMatcher in ranking more relevant services higher. Another influence factor may be that iMatcher applies machine learning techniques and has been trained with the binary relevance judgments of OWLS-TC. Switching to other definitions of relevance (e.g. strict binary relevance) seems to have a negative impact on iMatcher's performance relative to that of M3.

6 Conclusions and Future Work

In this paper, we investigated how to evaluate the retrieval effectiveness of SWS matchmakers based on graded instead of binary relevance. We discussed the notion of relevance in this particular context and proposed a well-founded scale of relevance levels customized to the SWS matchmaking domain. We presented a number of evaluation measures for graded relevance and described an experiment of using those measures to perform a retrieval evaluation of two SWS matchmakers.

We need to note once more, that our results have to be considered preliminary because of the nature of the test data used. First, there was a significant variation

in how the different judges judged our test data. We believe this to be caused by the insufficient documentation of the services in our data and expect this issue to improve if more realistic and better documented services are used than are currently available in the form of OWLS-TC. Second, we have compared only two matchmakers based on a relatively small data set. In terms of investigating the effects of different measures and different relevance scales it would be particularly desireable to have access to a larger number of directly comparable matchmakers. This will be the case if either more readily implemented matchmakers for a particular formalism (for instance SAWSDL) become available or test collections across formalisms will be developed.

Despite of these two restrictions, our results allow to draw a number of interesting conclusions. First, retrieval evaluation based on graded relevance is feasible both in terms of the effort to obtain graded instead of binary relevance judgments and in terms of the availability of measures suitable for graded relevance. Second, the choice of gain values (i.e. relevance levels) and the choice of measure has a significant influence on the evaluation results. Our results indicate that the choice of gain values has a greater impact than the choice of measure. Third, AWDP seems to be more sensitive towards changes in the parameterizations (regarding both, the penalty for late retrieval and changes of the gain values) than Q-measure and thus should probably be the first choice for future evaluations.

In our future work we plan to verify these findings with better data. As a first step, we would like to investigate whether relevance judgments really become more consistent across judges when more realistic and well documented services are used. A second step will then be to compare a larger number of matchmakers based on that more realistic data.

Acknowledgments

We would like to thank Patrick Kapahnke and Matthias Klusch for providing us with test data from the S3 Contest and for their help in resolving some technical problems with that data. We would also like to thank them and the other contributors for developing OWLS-TC and making it publicly available. Finally, we would like to thank Vassileios Tsetsos for giving us the graded relevance judgments from [3] to compare them with ours.

References

1. McIlraith, S.A., Son, T.C., Zeng, H.: Semantic web services. IEEE Intelligent Systems **16**(2) (2001) 46–53
2. Klusch, M.: Semantic web service coordination. In M. Schumacher, H.H., ed.: CASCOM - Intelligent Service Coordination in the Semantic Web. Springer (2008)
3. Tsetsos, V., Anagnostopoulos, C., Hadjiefthymiades, S.: On the evaluation of semantic web service matchmaking systems. In: 4th IEEE European Conference on Web Services (ECOWS2006), Zürich, Switzerland (2006)

4. Mea, V.D., Demartini, G., Gaspero, L.D., Mizzaro, S.: Measuring retrieval effectiveness with average distance measure (ADM). Information Wissenschaft und Praxis **57**(8) (2006) 405–416
5. Sakai, T.: New performance metrics based on multigrade relevance: Their application to question answering. In: Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization (NTCIR04), Tokyo, Japan (2004)
6. Noia, T.D., Sciascio, E.D., Donini, F.M.: Semantic matchmaking as non-monotonic reasoning: A description logic approach. Journal of Artificial Intelligence Research (JAIR) **29** (2007) 269–307
7. Kekäläinen, J., Järvelin, K.: Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology **53**(13) (2002) 1120–1129
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446
9. Sakai, T.: On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In: First International Workshop on Evaluating Information Access (EVIA), Tokyo, Japan (2007)
10. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. Information Processing and Management **43**(2) (2007) 531–548
11. Sakai, T.: Ranking the NTCIR systems based on multigrade relevance. In: Revised Selected Papers of the Asia Information Retrieval Symposium, Beijing, China (2004) 251–262
12. Kishida, K.: Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan (2005)
13. Mizzaro, S.: Relevance: The whole history. JASIS **48**(9) (1997) 810–832
14. Klusch, M., Fries, B., Sycara, K.: Automated semantic web service discovery with OWLS-MX. In: 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan (2006)
15. Kaufer, F., Klusch, M.: Performance of hybrid WSML service matching with WSMO-MX: Preliminary results. In: First International Joint Workshop SMR^2 on Service Matchmaking and Resource Retrieval in the Semantic Web at the 6th International Semantic Web Conference (ISWC2007), Busan, South Korea (2007)
16. Kiefer, C., Bernstein, A.: The creation and evaluation of iSPARQL strategies for matchmaking. In: 5th European Semantic Web Conference (ESWC2008), Tenerife, Canary Islands, Spain (2008) 463–477
17. Dudev, M., Kapahnke, P., Klusch, M., Misutka, J.: International semantic service selection contest s3 - contest participation guideline. online at http://www-ags.dfki.uni-sb.de/∼klusch/s3/s3-contest-plugin.zip (2007)
18. Küster, U., König-Ries, B.: On the empirical evaluation of semantic web service approaches: Towards common SWS test collections. In: 2nd IEEE International Conference on Semantic Computing (ICSC2008), Santa Clara, CA, USA (2008)
19. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.P.: Semantic matching of web services capabilities. In: First International Semantic Web Conference (ISWC2002), Sardinia, Italy (2002) 333–347
20. Preist, C.: A conceptual architecture for semantic web services (extended version). Technical Report HPL-2004-215, HP Laboratories Bristol (2004)
21. Keller, U., Lara, R., Lausen, H., Polleres, A., Fensel, D.: Automatic location of services. In: Second European Semantic Web Conference (ESWC2005), Heraklion, Crete, Greece (2005)