Measures for Benchmarking Semantic Web Service Matchmaking Correctness

Ulrich Küster and Birgitta König-Ries

Institute of Computer Science, Friedrich-Schiller-University Jena
D-07743 Jena, Germany
{Ulrich.Kuester|Birgitta.Koenig-Ries}@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 filtering and ranking a set of services with respect to a service query. Comparative evaluations of different approaches to this problem form the base for future progress in this area. Reliable evaluations require informed choices of evaluation measures and parameters. This paper establishes a solid foundation for such choices by providing a systematic discussion of the characteristics and behavior of various retrieval correctness measures in theory and through experimentation.

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. 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 filtering and ranking a set of services with respect to a service query. A variety of competing approaches to this problem has been proposed [1]. 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.

Evaluations in the area typically follow the approach taken in the evaluation of Information Retrieval (IR) systems: As a basis for the evaluation a test collection is provided. This collection contains a number of service offers, a (smaller) number of service requests and relevance judgments. These relevance judgments are provided by human experts and specify for each offer-request pair how relevant the offer is for the request, i.e., whether or to which degree the offer is able to satisfy the request. Matchmakers are then evaluated by comparing their output rankings with the one induced by the relevance judgments. This is done via retrieval correctness measures which assign an output ranking a performance score based upon the available reference relevance judgments.

While the general procedure is agreed upon, there has been little work up to now that investigates the influence of different settings on the stability and meaningfulness of the evaluation results. For instance, relevance judgments can be binary (relevant versus irrelevant) or graded (multiple levels of relevance), they may be subjective and different ways to deal with conflicting judgments are possible. Furthermore, a variety of evaluation measures with very different characteristics are available from IR.

Therefore, informed decisions about the evaluation measures employed, the underlying model of relevance and the procedure of how to obtain reliable relevance judgments are necessary for meaningful evaluations. In previous work we dealt extensively with the latter two issues [2]. We also presented a preliminary work discussing the applicability of different evaluation measures from IR to the SWS matchmaking domain [3]. In this paper, we extend this work by providing a comprehensive discussion of retrieval correctness measures in the domain of SWS matchmaking in theory and through experimentation.

The rest of the paper is structured as follows. In the following section, we provide an overview of related work in the area. Section 3 defines requirements to evaluation measures and discusses measures common in IR with respect to those requirements. A number of issues are identified and solutions to these issues are proposed. Section 4 complements this theoretic treatment by an analysis of the behavior of the measures in practice, based upon data from a community benchmarking event we organized. In particular we investigate the effects of three factors to the evaluation results: changes in the underlying definition of relevance, inconsistent relevance judgments and the choice of evaluation measure. The paper concludes with recommendations for appropriate decisions on evaluation measures that will make future evaluations more meaningful and reliable.

2 Related Work

Experimental evaluation of SWS retrieval correctness has received relatively little attention in the past [4]. Almost all approaches have so far relied on binary relevance and standard measures based on precision and recall without further motivating this evaluation approach.

Tsetsos et al. [5] were the first to raise the issue that binary relevance may be too coarse grained for reliable SWS retrieval evaluations. They proposed 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 of a service by an expert and a system under evaluation. However, a systematic investigation of the properties of different measures was not within the scope of their work. Apart from the work by Tsetsos et al. we are not aware of any work directly dealing with evaluation measures for SWS retrieval correctness evaluation.

In contrast, there is a large body of related work from the area of Information Retrieval that concerns the development and discussion of evaluation measures, e.g., [6–10]. However, it is not clear to which extent findings about stability

and sensitivity of measures from IR transfer to the domain of SWS retrieval, since there are important fundamental as well as practical differences between SWS retrieval and general IR evaluation [2]. Furthermore, we are not aware of a previous systematic discussion of the properties of all the measures covered in this paper, in particular not with respect to what we will define below as measure correctness.

Our work is directly related to the S3 Contest on Semantic Service Selection[1], an annual campaign dedicated to the comparative evaluation of SWS matchmakers. The most recent 2009 edition introduced the usage of some graded retrieval performance measures in addition to standard binary recall and precision and we organized the experiment we will use to analyze measure behavior in practice as part of this contest. By providing a systematic discussion of the characteristics and behaviors of all common retrieval measures, this paper aims at providing the foundation for well-founded choices of parameters and measures for future SWS retrieval correctness evaluations.

3 Retrieval Effectiveness Measures

Service matchmakers in the context of this paper compare a service request with a set of available service offers and return a list of matching services, ordered by decreasing estimated relevance to the request. Retrieval effectiveness measures need to quantify the quality of the output lists produced by various matchmakers. The following definitions will be used throughout this paper.

Definition 1 (Ranking). *A* ranking r *of a set of services* S *is an ordered sequence of the elements from* S*, i.e.:* $r = (r_1, r_2, \ldots, r_n), \quad n <= |S|, \quad r_i \in S, \quad r_i = r_j \Rightarrow i = j$*. The number* i *is called the* rank *of the service* r_i *with respect to the ranking* r*. A ranking with* $n = |S|$ *is called a* full ranking*.*

Definition 2 (Gain). *The* gain g *(*$g >= 0$*) of a service* s *with respect to a query* q *denotes the relevance of* s *to* q*. The function* g_q *which assigns each service* s *from a ranking* r *a gain* g *with respect to a query* q *is called a* gain function*. We furthermore define a binary flag that denotes whether a service at a given rank* i *is relevant or not:* $isrel_{r,q}(i) = 1$*, if* $g_q(r_i) > 0$ *and* 0 *otherwise.*

For the sake of simplicity, we will generally omit the query index q and the ranking index r in the following if the query or ranking under consideration is clear from the context or no specific query or ranking is referenced.

Definition 3 (Ideal ranking). *A full ranking* r *is called* ideal *iff it lists the services in decreasing order of relevance, i.e.:* $\forall i \in \{2..|S|\} : g(r_i) <= g(r_{i-1})$*.*

Definition 4 (Retrieval effectiveness measure). *A* retrieval effectiveness measure m *is a function which assigns a ranking* r *a value from* $[0, 1]$ *with respect to a gain function* g*:* $m_g(r) \rightarrow [0, 1]$*.*

[1] `http://dfki.de/~klusch/s3/`

Having introduced a basic notion of retrieval effectiveness measure, we now turn to defining desirable properties of such measures. Again, a few definitions are helpful.

Definition 5 (Ranking superiority). *A ranking* r *is called* superior *to a different ranking* r′ *with respect to a given gain function* g *(*$r > r'$*) iff* r′ *can be changed into* r *by a sequence of pair wise item swaps within* r′ *and for each two swapped items* r_i *and* r_j *from* r′ *it holds:* $i < j \Rightarrow g(r_i) < g(r_j)$ *(items with higher relevance are moved upwards).*

Definition 6 (Measure correctness). *A retrieval effectiveness measure m is called* correct *iff for any two rankings* r *and* r′ *and a gain function* g, $r > r' \Rightarrow m_g(r) > m_g(r')$ *holds.*

Ranking superiority and measure correctness formalize the intuitive notion that a ranking that lists items of higher relevance at comparatively higher ranks should always receive a superior effectiveness measure score. Besides this notion of correctness, three more properties of retrieval measures are desirable.

First, performance measures should allow to be compared meaningfully over queries. To avoid normalization problems, we require that an ideal ranking always receives a performance score of 1. Second, for graded relevance, measures should allow to configure the extent to which an item of comparatively higher relevance is preferred over an item of comparatively lower relevance. Third, for typical retrieval tasks, performance at the beginning of the output ranking is more important than performance at the end of the output ranking since a user typically will not completely read through a long ranking till its end. A good retrieval measure should thus emphasize top rank performance over bottom rank performance and allow to configure the extent of this emphasis.

3.1 Retrieval Measures from IR

After having briefly discussed desirable properties of retrieval effectiveness measures, we now turn to recalling some well established measures from IR. A complete coverage is beyond the scope of this paper, but available in [11, 12].

IR retrieval effectiveness measures are almost exclusively based upon the well-known *Recall* and *Precision* measures. Let R be the set of relevant items for a query and L be the set of the first l items returned in response to that query. Then $Recall_l$ is defined as the proportion of all relevant items that are contained in L and $Precision_l$ as the proportion of items in L that are relevant:

$$Recall_l = \frac{L \bigcap R}{R}, \; Precision_l = \frac{L \bigcap R}{L}.$$

Precision can then be measured as a function of Recall by scanning the output ranking from the top to the bottom and observing 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. If a system's performance needs to be captured in a single measure, the common one is *Average Precision* over relevant items:

$$AveP = \frac{1}{|R|} \sum_{i=1}^{|L|} isrel(i) \frac{\sum_{j=1}^{i} isrel(j)}{i}.$$

Historically, IR evaluation has primarily been based on binary relevance [11]. However, since about 2000, there is an increased interest in measures based on graded or continuous relevance [12, 9]. Various proposals have been made to generalize the Recall and Precision based measures from binary to graded relevance. We briefly recall the most common ones.

All of them are based on or can be expressed in terms of *Cumulated Gain* proposed by Järvelin and Kekäläinen [7]. Intuitively, Cumulated Gain at rank i measures the gain that a user receives by scanning the top i items in a ranked output list. More formally, the *Cumulated Gain* at rank i is defined as $CG(i) = \sum_{j=1}^{i} g(r_j)$. Moreover, the *Ideal Cumulated Gain* at rank i, $ICG(i)$, refers to the cumulated gain at rank r of an ideal ranking. This allows to define the *Normalized Cumulated Gain* at rank i as the retrieval performance relative to the optimal retrieval behavior: $NCG(i) = \frac{CG(i)}{ICG(i)}$.

Normalized Cumulated Gain allows a straightforward extension of AveP which has sometimes been referred to as *Average Weighted Precision* [6]:

$$AWP = \frac{1}{|R|} \sum_{i=1}^{|L|} isrel(i) \frac{CG(i)}{ICG(i)}.$$

Unfortunately, NCG(i) has a significant flaw that AWP inherits. ICG(i) has a fixed upper bound ($ICG(i) <= ICG(|R|)$). Thus, NCG(i) and AWP cannot penalize late retrieval of relevant items properly since NCG(i) cannot distinguish at which rank relevant documents are retrieved for ranks greater or equal than $|R|$ [6]. Several measures have been proposed that resolve this flaw of AWP.

Järvelin and Kekäläinen [7] 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 i as $DCG(i) = \sum_{j=1}^{i} \frac{g(i)}{disc(i)}$ with $disc(i) >= 1$ being an appropriate discount function. Järvelin and Kekäläinen suggested to use the log function and use its base b to customize the discount which leads to

$$DCG_{\log_b}(i) = \sum_{j=1}^{i} \frac{g(i)}{max(1, \log_b i)}.$$

An according definition of *Ideal Discounted Cumulated Gain* ($IDCG(r)$) can be used to define the *Normalized Discounted Cumulated Gain* at some document cutoff level l ($NDCG_l$) and a straightforward Version of AWP that we call *Average Weighted Discounted Precision*: $AWDP = \frac{1}{|R|} \sum_{i=1}^{|L|} isrel(i) \frac{DCG(i)}{IDCG(i)}$.

Kishida [9] proposed a generalization of AveP that also avoids the flaw of AWP:

$$GenAveP = \frac{\sum_{i=1}^{|L|} isrel(i)\frac{CG(i)}{i}}{\sum_{i=1}^{|R|} \frac{ICG(i)}{i}}.$$

Sakai [6] 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:

$$Q\text{-}measure = \frac{1}{|R|}\sum_{i=1}^{|L|} isrel(i)\frac{\beta CG(i) + \sum_{j=1}^{i} isrel(j)}{\beta ICG(i) + i}.$$

Finally, it has also been proposed to use Kendall's τ or other rank correlation measures to measure retrieval effectiveness by comparing a ranking r with an ideal ranking r' [13]. Kendall's τ measures the correlation between two rankings via the number of pair wise adjacent item swaps that are necessary to turn one ranking into another. Since Kendall's τ yields values between 1 (identical rankings) and -1 (inverse rankings), it needs to be normalized to yield values from $[0, 1]$: $\tau'(r) = \frac{\tau(r,r')+1}{2}$.

3.2 Discussion of Measures

With the exception of τ' and AveP, all measures introduced above allow fine-tuning the extent to which highly relevant items are preferred over less relevant items by choosing an appropriate gain function. Furthermore, except for CG_l and DCG_l all measures are properly normalized and assign an ideal ranking a score of 1. We now discuss the measures with respect to correctness and the degree of control over the extent to which late retrieval is penalized. For illustration, please consider the rankings displayed on the left side in Table 1. The numbers in the rankings represent gain values or corresponding items to be retrieved. The right side of the table provides the performance scores that various measures assign to the given rankings. Please observe that R_1 is the optimal ranking and that $R_1 > \{R_2, R_3\} > R_4 > R_5 > R_6 > R_7$. Furthermore, R_2 should be considered preferable to R_3 since the single item swap compared to the optimal ranking occurs at lower ranks than is the case with R_2. These relations should be reflected in the performance scores assigned by the measures. This is not always the case as will be discussed below.

AveP: Trivially, binary AveP can not distinguish among items of different relevance grades and is thus not correct for graded relevance: $AveP(R_1) = AveP(R_2)$.

NDCG: NDCG_l is correct, if the used discount function is valid, i.e., positive and strictly monotonic increasing for $i \in [1, l]$. Notably, this is not the case for the originally suggested and typically used $max(1, \log_b(i))$ discount function which is constant for ranks 1 through b. With valid discounting functions (e.g., $\sqrt{i}$), however, NDCG_l is correct as far as rankings are only considered up to rank l.

$R_1 = (10, 6, 3, 0, 0, 0, 0, 0, 0, 0)$
$R_2 = (10, 3, 6, 0, 0, 0, 0, 0, 0, 0)$
$R_3 = (6, 10, 3, 0, 0, 0, 0, 0, 0, 0)$
$R_4 = (3, 6, 10, 0, 0, 0, 0, 0, 0, 0)$
$R_5 = (0, 0, 0, 3, 6, 10, 0, 0, 0)$
$R_6 = (0, 0, 0, 0, 0, 10, 6, 3, 0)$
$R_7 = (0, 0, 0, 0, 0, 0, 10, 6, 3)$

	R_1	R_2	R_3	R_4	R_5	R_6	R_7
$AveP$	1.00	1.00	1.00	1.00	0.38	0.28	0.24
$NDCG_9(\sqrt{i})$	1.00	0.98	0.93	0.81	0.52	0.46	0.43
AWP	1.00	0.94	0.87	0.62	0.54	0.79	0.79
$Q\text{-}measure(\beta = 1)$	1.00	0.94	0.88	0.66	0.50	0.65	0.63
$GenAveP$	1.00	0.94	0.84	0.57	0.23	0.26	0.23
$AWDP(\sqrt{i})$	1.00	0.94	0.81	0.54	0.29	0.37	0.35
τ'	1.00	0.97	0.97	0.92	0.67	0.58	0.50

Table 1. Comparison of evaluation measures

NDCG_l also allows configuring the extent to which late retrieval is penalized by choosing a more or less quickly growing discount function.

AWP: As mentioned above, AWP can not differentiate among rankings that are equal till rank $|R|$, e.g., $AWP(R_6) = AWP(R_7)$. Even worse, the order of items may matter more than their absolute rank, e.g., $AWP(R_5) < AWP(R_6)$, despite of $R_5 > R_6$. AWP is thus not correct. To the best of our knowledge, this order versus rank defect has not been discussed so far. AWP also does not allow configuring the extent to which late retrieval is penalized.

Q-Measure: Q-Measure was designed to resolve the first defect of AWP, but unfortunately inherits the second one, e.g., $Q\text{-}measure(R_5) < Q\text{-}measure(R_6)$. The actual vulnerability of Q-Measure to this defect depends upon the actual choices for the gain values and its β factor. But for any setting, it either inherits the vulnerability from AveP of not properly distinguishing among items of varying relevance or the order versus rank defect from AWP and is thus not correct. Q-Measure provides limited control over the extent to which late retrieval is penalized via its β factor.

GenAveP: GenAveP shares the order versus rank defect with Q-Measure and AWP, e.g., $GenAveP(R_5) < GenAveP(R_6)$. Therefore, just like Q-Measure and AWP, it is not correct. However, in practice, GenAveP seems to be somewhat less vulnerable to the mentioned defects than the other two measures. GenAveP does not allow configuring the extent to which late retrieval is penalized.

AWDP: AWDP resolves the first defect of AWP if the used discounting function is valid. Nevertheless it inherits the order versus rank defect from AWP, e.g., $AWDP_{\sqrt{i}}(R_5) < AWDP_{\sqrt{i}}(R_6)$. It is therefore also not correct. Like the choice of β for Q-Measure, the choice of a discount function for AWDP has an influence on its practical vulnerability to this particular defect. By choosing a proper discount function AWDP allows configuring the extent to which late retrieval is penalized.

Rank Correlation Measures: Kendall's τ, respectively τ', is correct in the sense provided above. However, it does not differentiate between swaps that occur at

the top and those that occur at the bottom of a ranking, e.g., $\tau(R_2) = \tau(R_3)$. Furthermore, as mentioned above, it also does not allow to configure the extent to which highly relevant items are preferred over less relevant ones.

Summary: It is remarkable, that, as can be seen from this discussion, with the exception of NDCG and Kendall's τ all commonly used evaluation measures based on graded relevance are not correct in the sense defined above. Furthermore, NDCG is typically used with a discount function that renders it effectively incorrect, too, and Kendall's τ lacks the ability of emphasizing top versus bottom rank performance and configuring the extent to which highly relevant items are preferred over marginally relevant ones.

3.3 Proposed Improvements

After having discussed shortcomings of most commonly used measures for graded relevance, we now propose improvements to avoid these shortcomings. Table 2 shows a comparison of the original with the altered versions of the measures that illustrates how the altered versions avoid the problems of the original ones: in contrast to the scores of AWP, GenAveP and AWDP, those of ANCG, GenAveP′ and ANDCG are strictly decreasing from R_1 to R_7.

NDCG: The issues with NDCG can be trivially avoided by using an adapted version of the original discount function, namely $disc(i) = \log_b(i + b - 1)$, or any other valid function, like a root function, i.e., $disc(i) = i^a, 0 < a <= 1$. Such obvious adaptations have been proposed previously, e.g., [14]. Therefore, it is somewhat surprising to see that most literature still uses the original flawed discounting functions, e.g., [8].

AWP/AWDP and GenAveP: The defects of AWP/AWDP and GenAveP can be avoided by not averaging over relevant items only, but over all items, i.e.:

$$AW(D)P' = \frac{1}{|R|} \sum_{i=1}^{|L|} \frac{(D)CG(i)}{I(D)CG(i)}, \quad GenAveP' = \frac{\sum_{i=1}^{|L|} \frac{CG(i)}{i}}{\sum_{i=1}^{|R|} \frac{ICG(i)}{i}}.$$

AW(D)P' can be interpreted as the area under a N(D)CG-chart [7]. To properly distinguish the altered from the original versions, we will refer to the altered ones as *Averaged Normalized Cumulated Gain (ANCG)* and *Averaged Normalized Discounted Cumulated Gain (ANDCG)* in the following.

Others: In contrast to the previous measures, Q-Measure can not be fixed in the same fashion. Averaging over all, and not only relevant items, decreases the performance value of the AveP part of Q-Measure to values much smaller than 1.0 even for optimal rankings if the number of relevant items is much smaller than the total number of items. This makes averaging of results over queries with differing numbers of relevant items unstable.

	R_1	R_2	R_3	R_4	R_5	R_6	R_7
AWP	1.00	0.94	0.87	0.62	0.54	0.79	0.79
$ANCG$	1.00	0.98	0.96	0.87	0.51	0.37	0.26
$GenAveP$	1.00	0.94	0.84	0.57	0.23	0.26	0.23
$GenAveP'$	1.00	0.97	0.91	0.76	0.30	0.20	0.13
$AWDP$, $disc(i) = \sqrt{i}$	1.00	0.94	0.81	0.54	0.29	0.37	0.35
$ANDCG$, $disc(i) = \sqrt{i}$	1.00	0.96	0.89	0.72	0.27	0.18	0.12

Table 2. Comparison of altered evaluation measures

Similarly the issues with Kendall's τ can also not been fixed easily. Rank correlation measures are not designed to distinguish between whether rankings differ at the top or bottom. Furthermore, rank correlation does not offer an intuitive way of configuring the extent to which highly relevant items are preferred over less relevant items.

3.4 Conclusions

The discussion above has shown that various measures for graded relevance are available, but that even some of the common ones behave unintuitively in certain cases. A fix for the problems associated with AWP, AWDP and GenAveP has been proposed. With this fix, NDCG, ANCG (fixed AWP), ANDCG (fixed AWDP) and GenAveP' (fixed GenAveP) are correct as defined above.

While this correctness guarantees a ranking of matchmakers which corresponds to intuition if the matchmaker's output rankings are pair wise superior, it does not guarantee a good ranking of matchmakers that produce outputs that are not pair wise superior, the common case in realistic settings. For such rankings, there is no objective notion of superiority, since a decision has to be made how to balance highly against less relevant items and performance in top against that in lower ranks (or recall versus precision for that matter).

The following Section 4 will thus complement the already presented discussion by an investigation of the behavior of the covered measures based on real rankings in a realistic retrieval experiment. Based upon this investigation, recommendations for retrieval effectiveness measures will be provided in Section 5.

4 Analysis of Measure Behavior in Practice

We organized an evaluation of semantic service matchmakers across formalisms as part of the 2009 S3 Contest on Semantic Service Selection. Full information about this evaluation campaign, its setup and results is available online[2]. Due to space restrictions, we will only provide a brief introduction to the setup of the evaluation before discussing the characteristics of retrieval correctness measures based upon the data gathered from the evaluation.

[2] `http://fusion.cs.uni-jena.de/professur/jgdeval`

Goals: The evaluation targets the use case of a human developer that is searching for a Web service that provides a functionality needed in some application being developed. Semantic service matchmakers are expected to make this discovery process more efficient by providing efficient filtering and ranking of services in registries. The task being evaluated is thus to rank a list of given Web services with respect to their relevance to given user queries. Relevance is defined by reference judgments from human experts (see evaluation parameters below).

Data Set: For the evaluation, a data set of real services with rich information was needed. Furthermore, to make the retrieval task challenging, a large number of related by slightly different services was desired. Existing data sets did not meet these requirements in an ideal way [15]. Thus, the Jena Geography Dataset (JGD) was created[3] [2]. This data set consists of 200 real service operations from the geography domain which have been collected with all the information available online, i.e. all the information that a human developer finds when searching these services.

Experimental Setup: The experiment was executed in multiple phases. In the first phase, services were released to participating groups and the particants provided (semantic) annotations for the services in a way that they felt most suitable for their needs and matchmakers. Unfortunately, participants were overcharged by annotating the full 200 services and the dataset had to be reduced to 50 services. In a second phase, nine requests were released. Relevance judgments were not released together with the requests and the participating groups were asked to have members formalize the queries who had not been involved in the annotation of the services previously.

Finally, participants had to provide an implementation of their matchmaking system, pluggable to the SWS Matchmaker Evaluation Environment (SME2)[4] via a predefined interface. After services, queries and ontologies had been collected, the matchmakers were installed on a dedicated machine. SME2 was used to execute the evaluation, i.e., send queries to the registered matchmakers and retrieve the returned service rankings.

Five groups participated with six matchmakers (Themis-S, WSColab, IRS-III, SAWSDL-MX1/MX2, SAWSDL iMatcher) in the experiment, representing a variety of approaches from NL processing via folksonomy tagging to the usage of logic semantic annotations of services. Details on these matchmakers can be found online. Furthermore, we added the average performance of 50 random service rankings to the results as a performance bottom line.

Evaluation Parameters: Relevance judgments for the JGD have been created according to a multi-dimensional graded relevance scale which differentiates among the interface compatibility, the functional completeness and the functional equivalence of services. The judgments have been created by three judges independently. Afterwards, consensus judgments were built by debating judgments that

[3] `http://fusion.cs.uni-jena.de/professur/jgd`
[4] `http://projects.semwebcentral.org/projects/sme2/`

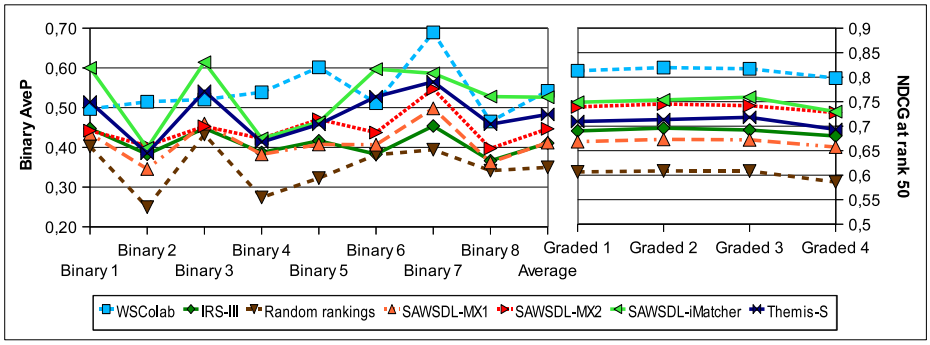

Fig. 1. Sensitivity of binary AveP and $NDCG_{50}$ to changes in the relevance definition

differed among judges. Again, full information is available online and in previous work [2]. The measures described in Section 3 allow evaluating SWS retrieval systems based on graded relevance but leave open the question about the proper parameter combinations to use in an evaluation. In order to investigate the effects of different definitions of relevance, we used four different gain value settings for the graded relevance and eight different definitions of binary relevance, i.e. different ways how to reduce the multi-dimensional graded relevance judgments to binary ones.

4.1 Influence of Relevance

We now turn to analyzing the characteristics of the discussed retrieval correctness measures and start with the effect of changes in the relevance definition to the evaluation results. We concentrate on the question whether a measure correctly orders the matchmakers by their retrieval effectiveness and is not influenced by other factors not of interest and under control during the evaluation.

Figure 1 illustrates the sensitivity (changes in the relative order of evaluated matchmakers) of binary AveP and $NDCG_{50}$ with discount $\log_2(i+1)$ to changes in the relevance definition. It highlights drastic swaps in the relative performance of the evaluated matchmakers if binary relevance is used (left side). The usage of Binary 2 compared to Binary 3, for instance, results in largely different evaluation results. These findings are in line with similar studies from IR, e.g. [16].

In contrast, measures based on graded relevance are almost entirely stable against moderate changes in the gain values. Using $NDCG_{50}$ with discount $\log_2(i+1)$, for instance, there was not a single swap in matchmaker order for the four different graded relevance settings (right side). This finding is, again, in line with similar findings from the IR community [8]. Nevertheless the amount of difference in stability is remarkable. At least our test data makes a very strong case for preferring graded over binary relevance for the given evaluation use case.

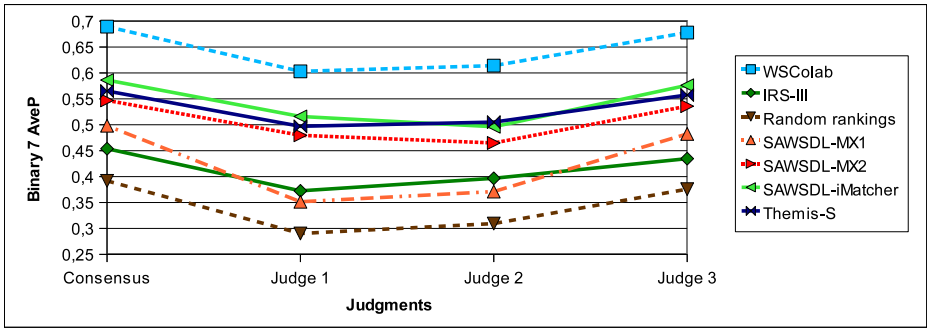

Fig. 2. Sensitivity of AveP to inconsistent relevance judgments (Binary 7 relevance)

4.2 Influence of Relevance Judge

It is well known from IR, that relevance judgments for retrieval evaluations differ among judges and for the same judge at different times [10]. In previous work, we investigated this issue in depth in the context of relevance judgments for service retrieval evaluation. We found significant inconsistency in judgments in this domain, too [2]. We are now able to complement the corresponding discussion by analyzing the effect that judgments by different judges really have on the comparative evaluation results.

Figure 2 show the computed AveP scores for the most liberal binary relevance setting using the consensus judgments as well as the original ones obtained from each of the three judges. The figure illustrates that changes in rankings, even notable ones, do occur but also that the influence is much smaller than that of switching the definition of relevance. Again, graded relevance (not shown in the figure) was more stable than binary relevance. However, with the exception of $NDCG_{50}$, swaps in rankings occurred occasionally using graded relevance, too.

4.3 Influence of Evaluation Measure

Finally, we now turn to discussing the influence of the choice of evaluation measure to the evaluation results. We consider NDCG, ANDCG, AWDP, ANCG, AWP, Q-Measure ($\beta \in \{0.5, 1, 2\}$), AveP, GenAveP and GenAveP'. The measures including a discount are analyzed using $\sqrt{i}$, $\log_2(i+1)$, $\log_3(i+2)$ and $\log_5(i+4)$ as discount function.

Figure 3 shows the performance scores from these measures using the Graded 1 relevance setting (AveP is computed assuming all services with a positive gain as relevant). The figure illustrates that the choice of evaluation measure influences evaluation results and also demonstrates the issues discussed in theory in Section 3.

As can be seen, there is a drastic difference in measure behavior between the incorrect AWP and its fixed counterpart ANCG. Please recall, that AWP had

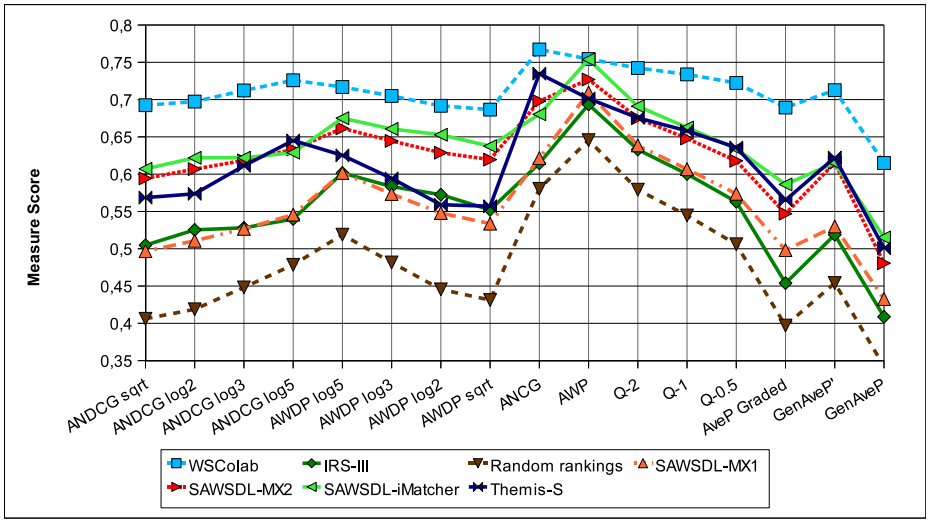

Fig. 3. Comparison of graded evaluation measures

two defects. First, its inability to punish very late retrieval, second, its property of rewarding correct order of relevant items rather than their absolute ranks. Themis-S is the only matchmaker whose score declines when switching from ANCG to AWP. This can be primarily explained through the first AWP defect, since Themis-S is inferior at retrieving highly relevant items at the top ranks, but superior at retrieving all relevant items relatively soon (see detailed evaluation results online). The first characteristic is correctly punished by both measures, whereas the second is not rewarded by AWP.

However, Themis-S is also evaluated comparatively poorly by the AWDP measures, which suffer from the order versus rank defect, but not from the inability of properly punishing late retrieval. This bias of AWDP against Themis-S is particularly evident by comparing AWDP and its correct counterpart ANDCG (please note that $NDCG_{50}$ rated equal to ANDCG and was thus not included in the chart). A comparison of scores within the different versions of ANDCG and AWDP illustrates nicely the effect of discounting. Stronger discounting comparatively benefits IRS-III whereas Themis-S profits from smaller discounts. This is an expected behavior and results from IRS-III performing a precision oriented matchmaking versus the recall oriented matchmaking of Themis-S. The fact that stronger discounting penalizes Themis-S is another argument for the bias of AWDP against Themis-S being caused by the order versus rank defect and not an insufficient punishment of late retrieval.

It is notable, that Q-Measure and GenAveP, which also suffer from the order versus rank defect, are less biased against Themis-S than AWDP. Still, the fixed GenAveP' and Q-Measure with a small β are more favorable for Themis-S than the incorrect GenAveP and Q-Measure with a larger β.

5 Summary and Conclusions

This paper dealt with measures for evaluating the retrieval correctness of SWS matchmakers. To the best of our knowledge it is the first such work in the area of SWS retrieval. Desireable properties of evaluation measures were defined and various measures from IR introduced. Properties of these measures were first discussed in theory. Defects in some measures were identified and fixes for these defects proposed. Finally, the theoretic discussion was complemented by an experimental investigation of measure behavior in practice. From the discussion and experimental analysis, some important conclusions for future retrieval effectiveness evaluations may be derived.

First, binary AveP is highly sensitive towards changes in the definition of relevance underlying the relevance judgments. Unless one knows about this definition very well, is certain that the definition matches the use case of the evaluation and that the reference judges applied the definition correctly, we recommend against using binary relevance in the future. In contrast, graded relevance is extremely stable against moderate changes in the gain values (and thus the underlying definition of relevance) and therefore should be preferred over binary relevance.

Second, inconsistency in relevance judgments influences evaluation results, but only moderately. Again, binary relevance is less stable than graded relevance. Obviously, more reliable judgments are preferable, but the effects of inconsistency seem to remain in a tolerable range, at least for graded relevance.

Third, the choice of evaluation measure influences the evaluation results. The choice of a graded measure has less influence than the choice of relevance with binary AveP, but more influence than inconsistent judgments. To obtain reliable evaluation results, one should not choose a particular measure without justifying the choice. For a fair and unbiased treatment, analysis with different measures and corresponding reporting is recommended. Contradicting measures indicate differing retrieval characteristics of the matchmakers exchanging ranks and thus allow tracing those characteristics. Corresponding insights are an important additional advantage of using different evaluation measures.

Fourth, as was suggested before, AWP is not a reliable evaluation measure because of its inability to properly punish late retrieval. However, Q-Measure, GenAveP and in particular AWDP may also show an unintuitive measure behavior. The alternative NDCG_l and the newly proposed ANCG/ANDCG are correct with respect to the definition provided in Section 3 and offer the most intuitive and flexible way of customizing the emphasis on top over bottom ranks. These measures are recommended for future retrieval effectiveness evaluations. NDCG charts are probably the most informative way of presenting evaluation results, since they provide an indication of the performance of matchmakers over ranks and still provide a summary measure by the value at the bottom rank (NDCG_{50} in our case). Attention should be paid to choosing a valid discount function for this measure.

We hope that these findings will help establishing sound evaluation methodologies and further advancing the state of the art in SWS matchmaking.

Acknowledgments: We owe great thanks to Patrick Kapahnke and Matthias Klusch from DFKI Saarbrücken for providing the SME2 tool and performing the actual execution of the evaluation on their machines. Additionally we would like to thank all participants in Track 3 of the 2009 S3 Contest for all their efforts without which this work would not have been possible.

References

1. Klusch, M.: Semantic web service coordination. In M. Schumacher, H.H., ed.: CASCOM - Intelligent Service Coordination in the Semantic Web. Springer (2008)
2. Küster, U., König-Ries, B.: Relevance judgments for web services retrieval - a methodology and test collection for sws discovery evaluation. In: Proc. of the 7th IEEE European Conference on Web Services (ECOWS09), Eindhoven, The Netherlands (2009)
3. Küster, U., König-Ries, B.: Evaluating semantic web service matchmaking effectiveness based on graded relevance. In: Proc. of the 2nd International Workshop SMR^2 on Service Matchmaking and Resource Retrieval in the Semantic Web at ISWC08, Karlsruhe, Germany (2008)
4. Küster, U., König-Ries, B., Petrie, C., Klusch, M.: On the evaluation of semantic web service frameworks. International Journal On Semantic Web and Information Systems **4**(4) (2008)
5. 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)
6. Sakai, T.: Ranking the NTCIR systems based on multigrade relevance. In: Revised Selected Papers of the Asia IR Symposium, Beijing, China (2004) 251–262
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446
8. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. Information Processing and Management **43**(2) (2007) 531–548
9. 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)
10. Saracevic, T.: Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. Library Trends **56**(4) (2008) 763–783
11. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
12. Demartini, G., Mizzaro, S.: A classification of IR effectiveness metrics. In: Proc. of the 28th European Conference on IR Research (ECIR06), London, UK (2006)
13. Melucci, M.: On rank correlation in information retrieval evaluation. ACM SIGIR Forum **41**(1) (2007) 18–33
14. Burges, C.J.C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.N.: Learning to rank using gradient descent. In: Proc. of the Twenty-Second International Conference on Machine Learning (ICML05), Bonn, Germany (2005) 89–96
15. Küster, U., König-Ries, B.: Towards standard test collections for the empirical evaluation of semantic web service approaches. International Journal of Semantic Computing **2**(3) (2008) 381–402
16. Voorhees, E.M.: Evaluation by highly relevant documents. In: Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR01), New Orleans, LA, USA (2001) 74–82