Inf Retrieval J (2016) 19:284–312
DOI 10.1007/s10791-015-9275-x

Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation

Thomas Demeester[1] · Robin Aly[2] · Djoerd Hiemstra[2] · Dong Nguyen[2] · Chris Develder[1]

Received: 1 May 2015 / Accepted: 15 October 2015 / Published online: 13 November 2015

Abstract Evaluation of search engines relies on assessments of search results for selected test queries, from which we would ideally like to draw conclusions in terms of relevance of the results for general (e.g., future, unknown) users. In practice however, most evaluation scenarios only allow us to conclusively determine the relevance towards the particular assessor that provided the judgments. A factor that cannot be ignored when extending conclusions made from assessors towards users, is the possible disagreement on relevance, assuming that a single gold truth label does not exist. This paper presents and analyzes the predicted relevance model (PRM), which allows predicting a particular result's relevance for a random user, based on an observed assessment and knowledge on the average disagreement between assessors. With the PRM, existing evaluation metrics designed to measure binary assessor relevance, can be transformed into more robust and effectively graded measures that evaluate relevance towards a random user. It also leads to a principled way of quantifying multiple graded or categorical relevance levels for use as gains in established graded relevance measures, such as normalized discounted cumulative gain, which nowadays often use heuristic and data-independent gain values. Given a set of test topics with graded relevance judgments, the PRM allows evaluating systems on different scenarios, such as their capability of retrieving top results, or how well they are able to

✉ Thomas Demeester
thomas.demeester@intec.ugent.be

Robin Aly
r.aly@utwente.nl

Djoerd Hiemstra
d.hiemstra@utwente.nl

Dong Nguyen
d.nguyen@utwente.nl

Chris Develder
cdvelder@intec.ugent.be

[1] Ghent University - iMinds, Ghent, Belgium

[2] University of Twente, Enschede, The Netherlands

filter out non-relevant ones. Its use in actual evaluation scenarios is illustrated on several information retrieval test collections.

Keywords Information retrieval evaluation · Test collections · Graded relevance assessments for information retrieval · Assessor disagreement

1 Introduction

Measuring the effectiveness of search results for users is essential for the improvement, comparison and tuning of search engines. To achieve this task, effectiveness measures often employ relevance labels assigned by assessors as ground truth. Hence, the literature often treats assessors as if they are actual users that pose a query on their current information need, and assess the returned results accordingly. However, in practice assessors are often workers with the task to assess the relevance of results for users they have never met. To estimate the impact of this assumption, previous work studies the disagreement of assessors on binary labels and its influence on search engine comparisons (Voorhees 2001), leading to the conclusion that search engine comparisons are stable even under substantial assessor disagreement. Demeester et al. (2014) show that in a graded relevance setting, this disagreement is especially strong on the top relevance levels. The current paper explicitly models the disagreement between assessors and particular scenarios of user relevance, such as users that are only satisfied with top results, or users that are looking for any result that is at least marginally relevant. This model is applied to graded relevance based evaluation. The findings are supported by experimental results on two different datasets.

Modeling the plurality of users involved in search currently receives much research interest. Most work focuses on differences among users, e.g., diversity of search results (Zhai et al. 2003) and query ambiguity (Agrawal et al. 2009). We propose that the plurality between users and assessors is equally important. So far, research on this topic considers deviations between assessors and users as mistakes, e.g., due to input error or ambiguous instructions, and evaluation measures had to prove to be stable against these unwanted effects. In this paper we consider differences between assessors' predictions and users as natural and we propose methods to integrate them into effectiveness measures.

We focus on exploiting our model for user–assessor plurality in standard evaluation measures based on graded assessment levels. Other measures, e.g., for ambiguity, could also benefit from our model. However, their particular consideration of differences among users deviates from the presented approach, which makes the connection between assessor observations and user preferences difficult to isolate. In a standard graded relevance evaluation setup, the gains for the relevance grades often lack a direct connection to the assumed evaluation scenario and are typically set heuristically. For example, Kanoulas and Aslam (2009) set gain values of the normalized discounted cumulative gain (nDCG) effectiveness measure by optimizing formal quality criteria for test topics, but they do not model the connection with relevance towards users. The graded average precision (GAP) by Robertson et al. (2010) is one of the first to define gain values based on users: they consider user populations that perceive documents as relevant from specific thresholds in the ground truth relevance levels onwards. However, setting the GAP gains requires the hard task of determining the distribution of threshold values over the population. Our model also assumes a binary notion of relevance for individual users. However, it avoids

the common assumption that there is a single ground truth, given by the labels assigned by the assessors.

The main contribution of this paper is the predicted relevance model (PRM), which captures differences between assessor judgments in order to estimate the relevance of documents for random users, with:

- an assessor model with multiple relevance levels,
- a user model based on binary relevance and linked to the assessor model, and
- a detailed estimation procedure of probabilities that quantify disagreement.

The PRM predicts the relevance for a random user, based on an observed assessment, and the expected relevance over different assessors. The model is built on the insights gained by Demeester et al. (2014), who introduced the User Disagreement Model (UDM). The main differences with respect to this previous work are: (1) refinements of the model, generalized with respect to only considering top relevance, (2) deeper analyses and insights into the model, (3) new insights into applying the model on binary and graded relevance measures, (4) an experimental analysis of using the model for evaluating retrieval systems, and (5) a validation based on two different IR evaluation collections. The differences between the PRM and the UDM are discussed in more detail in Sect. 3.4.

Note that Demeester et al. (2014) present evidence and a quantitative analysis of user disagreement, most of which will not be repeated here. For example, it was shown that for the FedWeb12 dataset (Nguyen et al. 2012), the inter-assessor disagreement was much stronger than the intra-assessor disagreement. The PRM does not explicitly model the intra-assessor disagreement. However, if the assessors lack consistency with their own judgments, this will also increase the level of disagreement between assessors, which is captured by the PRM.

We first provide an overview of related work in Sect. 2, focusing on assessment disagreement and graded relevance measures, before detailing the PRM in Sect. 3. We then present the datasets (Sect. 4) that we will use to quantitatively study the model. Section 5 explains how gains are set in the PRM, Sect. 6 analyzes the PRM parameters, and Sect. 7 presents retrieval system evaluation results. We conclude the paper and provide ideas for future research in Sect. 8.

2 Related work

The PRM introduced in this paper is related to evaluation approaches that investigate the plurality between users and assessors as well as effectiveness measures that use graded relevance levels. This section presents related work on these two aspects.

2.1 Modeling disagreement on relevance

Differences in relevance assessments for a particular result can originate from the actual difference in opinion by assessors, or from an error, e.g., due to ambiguous instructions. These phenomena are often jointly referred to as assessment disagreement.

Early works (e.g., by Harter 1996; Voorhees 2000; Sormunen 2002) study the influence of assessor errors on information retrieval evaluation using binary relevance judgments, and conclude that assessment disagreement has only minor effects on search engine comparisons. Based on these works, one could question whether extended models of

assessors, such as our PRM, can make a difference. However, Bailey et al. (2008) aggregate and compare these early works and demonstrate that assessors with a different task and domain expertise significantly affect search engine comparisons.

Furthermore, Vakkari and Sormunen (2004) study assessors[1] that reassess results judgments by TREC assessors in an interactive search scenario. They mainly observe disagreement between their assessors and the TREC assessors at a lower relevance level, with a better agreement on highly relevant results. The distribution of disagreement over the relevance levels does not impact the validity of our current work, i.e., the PRM remains valid for disagreement on lower levels (as in Vakkari and Sormunen 2004), and on top levels (as in Demeester et al. 2014).

Carterette and Soboroff (2010) also show that assessor disagreement has an effect on system comparisons for effectiveness measures considering graded relevance levels. Their identification of several prototypical assessor types (e.g., unenthusiastic, pessimistic, lazy) is particularly relevant for test collections that route away from trained and supervised judges towards poorly trained and autonomous judges, e.g., in crowd-sourcing contexts. They show how different types of assessor errors affect evaluation measures, and propose strategies to compensate for such errors, e.g., by reassessing certain results. Turpin et al. (2009) study the differences of using assessments that include summaries instead of only full documents on system comparison. They find that system effectiveness depends on the information that assessors have in order to make their decisions. Al-Harbi and Smucker (2014) investigate the difference in annotation behavior between so-called primary assessors, who create and judge test topics, and secondary assessors, who are paid to judge existing topics based on given query descriptions, and are less certain in their judgments.

Our work differs from these contributions because the PRM does not assume a single truth label for each document and query, and uses assessments solely as predictions of the relevance for an unknown future user. Also, the source of the disagreement does not impact the way it is modeled in the PRM and used for predicting relevance towards a random user. In an ideal scenario, assessment disagreement such as random annotation mistakes, which are complementary to the actual disagreement of assessors, should first be filtered from the assessments. However, the PRM is also suited to cope with these errors, and by modeling the resulting uncertainty on the assigned assessments ensures for a robust evaluation setting.

An alternative to the classical assessment of isolated results, are preference judgments, which lead to higher agreement levels (Carterette et al. 2008). Kazai et al. (2013) examine the relationship between assessor disagreement and click based measures, which more directly reflects web users. Trained assessors appear to have higher inter-assessor agreement and are more likely to agree with clicks. Their results suggest that pairwise judgments lead to more awareness of the possible intent, and therefore lowers disagreement. The approach of preference judgments is not further investigated in the current work.

Hosseini et al. (2012) concurrently model the relevance of documents and the accuracy of individual assessors, based on multiple labels per document. Compared to this work, our PRM instead uses a limited set of documents with double judgments to model the ability of average assessors in predicting the average relevance according to a well-defined notion of user relevance, and is applicable to documents with single judgments.

[1] Vakkari and Sormunen (2004) adopt the term ‘users’ for the persons reassessing documents. In our terminology, such persons are referred to as assessors.

Our PRM model mainly focuses on dealing with potential disagreement of users/assessors on the *relevance* of individual results. Besides in their judgment of relevance, users may also differ in the actions they take when browsing *ranked* results. For instance, Carterette et al. (2012) use click logs to compute posterior distributions for probabilistic models of user interactions and show that different "types" of user behavior exist, each of which may lead to a potentially different search system evaluation ranking. Metrics for that evaluation that are based on such more complex user models (which we leave out of scope for this paper on relevance disagreement) include rank-biased precision (RBP) (Moffat and Zobel 2008), expected reciprocal rank (ERR) (Chapelle et al. 2009), expected browser utility (EBU) (Yilmaz et al. 2010), and time-calibrated measures (Smucker and Clarke 2012). Such metrics mainly aim at appropriately accounting for the impact of the *rank* a result is placed at. The current paper rather focuses on setting the appropriate weight of a result depending on its *relevance*, as in graded relevance effectiveness measures, as discussed next. We note that Smucker and Clarke (2012) also integrate into their model the probability that a user considers a result relevant, by clicking or saving it, given that a NIST assessor judged it as relevant. In that aspect, there is a connection to the PRM approach.

2.2 Graded relevance effectiveness measures

Graded relevance effectiveness measures allow assessors to use more than binary relevance labels. According to Kanoulas and Aslam (2009), there are two main challenges with this class of effectiveness measures: (1) to set the gain for each relevance label, and (2) to define the discount of this gain according to the rank of the document. Järvelin and Kekäläinen (2002) propose the popular normalized discounted cumulative gain (nDCG) measure, which uses a heuristic to set the relevance level gains and a logarithmic discount per rank. However, Kanoulas and Aslam (2009) find that gains should be set according to a user model in order to ensure that the measure reflects real users. Zhou et al. (2014) propose to learn suitable gain and discount functions based on assessor preferences of rankings. Sakai (2007) compares 14 graded relevance measures with 10 traditional binary measures, and concludes that average nDCG at rank k (nDCG@k) is among the best effectiveness measures for graded relevance in terms of stability, sensitivity, and resemblance of system rankings, and is fairly robust with respect to the choice of gain values.

Chapelle et al. (2009) propose the Expected Reciprocal Rank (ERR) effectiveness measure, which measures the inverse expected effort required for a user to satisfy their information need, and assumes the knowledge of the probabilities R_i that the user is satisfied with document i. The discount function is therefore based on a user model but the work does not specify a user model to set R_i. This is achieved by the effectiveness measure GAP by Robertson et al. (2010) where each individual user is imagined to have a threshold label, above which they consider documents relevant. GAP then determines gains based on the distribution of users over these thresholds. However, empirically determining threshold distributions is hard and may depend on the query. Like GAP, our PRM also considers users that actually have a binary notion of relevance. Unlike GAP, the PRM employs the differences in the prediction of assessors to set gains. As it is often simpler to observe assessors than users, setting these gains is easier and requires less data. Furthermore, although not studied in this paper, it seems plausible that the parameters of our PRM can be used to arrive at the probabilities R_i of the ERR measure.

Voorhees (2001) studies the difference in effectiveness between using all relevant results and only highly relevant results. She finds that graded relevance measures are

unstable due to the low number of highly relevant documents. Our PRM model improves the stability of graded relevance effectiveness measures by smoothing the judgment of an assessor with the possibility that users disagree with this judgment. For example, it takes into account the probability that a random user may consider a result top despite a judgment below the top level.

The model presented in this paper leads to estimates of the gains for the various relevance levels according to the probability that a random user consider results assessed with these levels as relevant. Our model hence assumes a binary notion of user relevance. However, other choices are possible. Kekäläinen (2005) and Voorhees (2001) propose weighting of relevance grades based on the (speculated or heuristic) relative importance of the relevance levels to the users. Compared to binary relevance, this leads to a more complex notion of user relevance, and hence a more flexible evaluation scenario. The current work could be combined with these approaches. When adopting a more general (non-binary) model of user relevance, the probability of agreement with this user model, given an assessor judgment and based on the average disagreement, could be used to properly adapt the relevance gains. This however falls out of scope for the current paper.

3 The predicted relevance model

The PRM presented in this section formalizes and extends the ideas from our experimental investigation of user disagreement in Demeester et al. (2014), where also the original User Disagreement Model (UDM) was put forward. In the following we first formally define the PRM (Sect. 3.1). We then describe a first application of the PRM in counting relevant results (Sect. 3.2), which allows transforming binary evaluation measures into graded measures based on the probability of binary relevance for an average user (Sect. 3.3.1) and leads to an interpretation of using the nDCG measure with PRM-based gains (Sect. 3.3.2). Finally, the differences and advantages of the PRM with respect to the original UDM formulation are discussed (Sect. 3.4).

3.1 Definition of the PRM

In this section, we provide the definitions behind the PRM, each followed by the key ideas on (1) the distinction and link between users and assessors, (2) the quantification of disagreement, (3) the conditions for the validity of the PRM, (4) the construction of relevance gain values.

Definition 1 The considered user population of the IR system or search engine under evaluation, consists of individual users for whom a result is either relevant (R), or non-relevant to a query.

Definition 2 The assessors are part of the evaluation setup, and assign relevance labels to results, according to well-described graded (or categorical) assessment levels, indexed by $i = 0, \ldots, T$. The lowest level $i = 0$ represents non-relevance, and the highest level $i = T$ is defined as top relevance.

3.1.1 Users versus assessors

The distinction between *users* and *assessors* is essential to the PRM. The *user* model, on the one hand, corresponds to the classical binary notion of relevance for each individual user. Different users may have different opinions on the relevance of the same result. The *assessors*, on the other hand, are an essential part of the setup to evaluate retrieval systems. They assign different relevance grades according to how useful they predict a particular result to be to the users. A description of these relevance grades is part of the evaluation setup, and identical for all assessors. The assessor model corresponds with the typical scenario of graded relevance assessments for IR evaluation.

3.1.2 Intuition of the PRM

As will be described in the following sections, this setup allows evaluating how capable a system is in returning results relevant to a random user. The intuition behind it can be summarized as follows. A given result will be considered relevant by one user, while another might find it not relevant. The task of the assessor in an evaluation setup, usually implicitly amounts to try and assess how likely it is that a *random* user would consider the given result relevant. This is typically done using graded relevance levels. How informative the assessments are, depends on how well the assessors are able to put themselves in the position of a user, and on the average user disagreement. This intuition leads to the definition of parameters that quantify disagreement on relevance.

Definition 3 $p_{R|i} \triangleq$ the probability that a random user would consider a particular result relevant (R), given the knowledge of an independent assessor judgment with level i.

3.1.3 Disagreement parameters

With Definition 3, we model a particular result's relevance to a random user as a Bernoulli distributed variable. In fact, we model the user relevance of *any* result for which an assessment with level i was observed, as a Bernoulli variable with success rate $p_{R|i}$. The parameters $p_{R|i}$ are called the disagreement parameters, as they are subject to the disagreement between an assessor and a user.

3.1.4 Assessors judgments for predicting user relevance

In a practical evaluation setup, the disagreement between user and assessor will be modeled from observations between different assessors, as these are the only ones observed. Consequently, the model only allows making claims towards the user population if the assessors are capable of putting themselves in the position of the user, at least on average over their assessments. This condition is not obvious in practice. Primary assessors are more likely able to judge results from the perspective of users, whereas secondary assessors are known to make more uncertain decisions (Al-Harbi and Smucker 2014). The observed disagreement among secondary judges, or between primary and secondary judges, may therefore not reflect the disagreement with respect to users. Specific details on the assessors in our experimental setup will be given in Sect. 4.

3.1.5 Linking user model and assessor model

Even if the assessors can put themselves in the position of typical users, another condition needs to be fulfilled, in order to move from modeling disagreement between assessors in terms of assessment levels, to modeling binary relevance for users. We need to be able to map the assessment levels to the binary notion of user relevance. How this is done, depends on the goal of the evaluation, and the nature of the assessment levels. Although the PRM remains valid for any set of categorical relevance levels, for this paper we make the simplifying assumption that they are graded, and can be indexed from the lowest to the highest level $i = 0, \ldots, T$, as in Definition 2. In that case a logical choice to define the user model is by means of a threshold θ on the assessment levels. For levels on or above this threshold ($i \geq \theta$) the user model assumes relevance, and non-relevance below the threshold ($i < \theta$). We illustrate this with two examples.

In a web search evaluation scenario, the goal may be to evaluate a system's capability of retrieving top relevant results, for example for highly precision-oriented applications. In that case, we assume that users are only satisfied with top results. This means the threshold for user relevance is at the top graded relevance level ($\theta = T$). This was the choice made for the initial introduction of the UDM by Demeester et al. (2014).

In a more lenient evaluation scenario, typical for recall-oriented applications, the user relevance threshold could be chosen just above non-relevance ($\theta = 1$), to indicate that users are satisfied with any at least marginally relevant result, or to test a search engine's capability of filtering out non-relevant results. Defining user relevance in such a way allows adapting the evaluation strategy for different applications or types of users.

3.1.6 Asymptotic case without disagreement

In the asymptotic case of a perfectly controlled environment with deterministic annotation rules, and without any disagreement among assessors or users, we would find $p_{R|i \geq \theta} = 1$ and $p_{R|i < \theta} = 0$. Evaluation based on the PRM would boil down to classical binary evaluation at threshold θ.

3.1.7 Extension towards multiple random users

Definition 3 can be extended by considering the binomial distribution of relevance over multiple users, instead of the Bernoulli distributed relevance of a single user. For example, based on $p_{R|i}$, the probability can be calculated that at least M out of N random users would consider the result assessed with level i as relevant, as shown by Demeester et al. (2014). This allows rescaling the disagreement parameters in a consistent way, with a probabilistic interpretation, in order to adapt the evaluation setup towards a stricter or more lenient interpretation of relevance. This will not be pursued any further in the current work.

3.1.8 Modeling assessor behavior

A final point of discussion is on situations in which the assessors are not able to imitate the users' notion of relevance, and thus the PRM cannot make claims on the user population, as indicated before. Even then, using the PRM has clear advantages with respect to heuristics, although the evaluation scenario would only model the assessor behavior, not the actual user population. For example, in the particular case of noisy crowd-sourced

relevance judgments, it is doubtful that the assessors have a good understanding of what the users are like. However, the goal is still to use these assessments to evaluate IR systems. Given the large assessor disagreement, the probability $p_{R|i<T}$ would be quite high, or $p_{R|T}$ rather low, as confirmed by Demeester et al. (2014). An evaluation based on a heuristic choice of relevance gains, such as gains exponential in the relevance grade i, may rely too strongly on the top judgments, and lead to a questionable robustness with respect to the choice of assessors. The PRM gains are adapted to the disagreement, and prevent an incorrect resolution between the systems under evaluation if the assessor disagreement does not allow it. This is in line with the work from Smucker and Clarke (2012), who argue that metrics which fail to model user variance overestimate the effect size of differences between retrieval systems. For example, consider the extreme case that relevance grades are randomly assigned. When comparing retrieval systems based on these assessments, no valid conclusions can be made. A traditional evaluation setup would incorrectly favor IR systems that highly rank top judged documents, especially if based on a limited number of test topics. According to the PRM, however, no difference between any of these systems would be detected, because the gains for all relevance grades would have equal values.

3.1.9 Applicability of the PRM

We conclude by saying that the PRM is widely applicable, taking into account disagreement between assessors. Whether the results allow making conclusions about the user population, or only represent the assessors, depends on how well assessors are able to judge from the users' perspective. This holds in general when evaluating IR systems based on assessor judgments, just like the assumption that the judged search results and test topics are representative for how the systems will be used in practice.

3.2 Counting relevant results

Before showing how the PRM can be used to set relevance weights in existing evaluation measures, we consider the task of counting the number of relevant results N_R in a set of N results, each with an associated relevance assessment. Let n_i indicate the number of results assessed with level i (with $\sum_i n_i = N$). The link between the binary user model and the assessment grades is defined by a threshold θ, as described in the previous section. If we neglect any disagreement between assessors or users, and purely estimate the number of relevant results from the assessments, we find

$$N_R^{\text{bin}} = \sum_{i \geq \theta} n_i, \tag{1}$$

in which the superscript 'bin' indicates the binary model based on the assessments alone. Taking into account the disagreement, the PRM leads to

$$N_R^{\text{PRM}} = \sum_{i=0}^{T} n_i p_{R|i}. \tag{2}$$

Equation (2) is the summation for each relevance grade i of the expected values $n_i p_{R|i}$ of the binomially distributed number of relevant results, given an observed assessment with level i, in n_i trials. This leads to the interpretation of N_R^{PRM} as the total expected number of

relevant results in the results set for a random user. In the following section we show how this result can be used to interpret evaluation measures that make use of the PRM, and in Sect. 7.1, we will give an experimental illustration.

3.3 The PRM and evaluation measures

This section describes how the PRM can be applied to binary evaluation measures that are based on counts of relevant results (Sect. 3.3.1), and how the nDCG measure can be interpreted from the PRM perspective (Sect. 3.3.2).

3.3.1 Binary evaluation measures

There are a number of established binary evaluation measures that rely on the number of returned relevant results. For such measures, the assessed number of relevant results in a traditional binary setting can be replaced by the expected number of relevant results according to the PRM. For measures that are linear in the number of relevant results, this leads to the expected value of that binary measure for a random user, as opposed to the value for the assessor alone. For example, the expected precision at rank N based on the PRM would be N_R^{PRM}/N, with Eq. (2). This allows transforming a binary evaluation measure effectively into a graded measure, still measuring binary relevance for users, but whereby the weights of the relevance grades represent the uncertainty on the assessors' judgments with respect to user preferences.

3.3.2 Graded evaluation measures

A similar reasoning is also possible for graded relevance measures. We will discuss the case of the normalized discounted cumulative gain (nDCG) measure, given its popularity. We will thus also use nDCG for our experimental results in Sect. 7. The application of the PRM to other measures is left open for future research.

Given a ranked results list, the nDCG measure incorporates the relevance of the result at rank r by means of the gain $g(i(r))$ which is a function of the relevance level i of that result. The cumulative gain at rank k (CG@k) is defined as the sum of the gains for each result up to that rank. Typical gain values used in literature are the exponential gain $(2^{i(r)} - 1)$ or the linear gain $i(r)$. Assuming that results at higher ranks are less likely to be reached by the user, the discounting factors $c(r)$ are introduced, leading to the discounted cumulative gain at rank k, similar to Zhou et al. (2014), as

$$\mathrm{DCG@k} = \sum_{\mathrm{r}=1}^{\mathrm{k}} \mathrm{c(r)g(i(r))}. \tag{3}$$

The discount factors used most often in literature are the logarithmic discount $c(r) = 1/\log(r+1)$, in which the gain a user obtains by moving down a ranked list drops less sharply than with the Zipfian discount $c(r) = 1/r$ (Kanoulas and Aslam 2009). The nDCG@k measure is obtained by normalizing DCG@k calculated from the ranked list of retrieved results, by the ideal DCG@k when based on a perfect ranking, i.e., according to decreasing relevance levels.

We propose to calculate the nDCG@k measure with the PRM disagreement parameters as gains, $g^{\mathrm{PRM}}(i) = p_{R|i}$, in order to model the relevance towards an average user. The choice of discount factors remains open, as the PRM is not suited to model the rank-

dependence of relevance in a results list. For our experiments, we use the logarithmic discount function. Our proposal for using the disagreement parameters as relevance gains can be motivated as follows, in a similar way as in Sect. 3.3.1, i.e., by considering the binary relevance perspective for a random user.

We assume the binary notion of user relevance introduced in Sect. 3.1, based on a threshold θ on the relevance grades. The corresponding binary gain values can be defined as $g^{\text{bin}}(i) = 1$ if $i \geq \theta$, and $g^{\text{bin}}(i) = 0$ otherwise. The cumulative gain CG@k based on $g^{\text{bin}}(i)$ can be interpreted as the number of relevant results among the top k retrieved results purely based on the assessor, ignoring any disagreement. The binary DCG@k, according to Eq. 3 but based on $g^{\text{bin}}(i)$, reduces to summing the discount factors of those ranks ($r \leq k$) with a result on or above the threshold ($i(r) \geq \theta$). The normalization factor for the binary nDCG@k is calculated as the binary DCG@k for the ideal ranking that places all results with grade $i \geq \theta$ before the others.

Using the PRM gains $g^{\text{PRM}}(i)$ leads to the interpretation of the resulting CG@k as the expected number of relevant results up to rank k, and of the resulting DCG@k as the expected value of the binary DCG@k, for a random user. The ideal ranking needed for the normalization in nDCG@k is based on decreasing relevance gains, in other words, based on the decreasing probability of relevance to a random user, given the assessor label.

With this approach, no ad-hoc quantification of the relevance level gains is needed. The relevance gains emerge naturally as the PRM disagreement parameters when calculating the expected value of the binary DCG measure for a random user.

3.4 Advantages of the PRM versus the UDM

This section explains the differences between the PRM and the UDM, focusing on the differences between the respective user models.

3.4.1 The UDM user model

The UDM introduced by Demeester et al. (2014) is based on a different user model, compared to the PRM presented in the current paper. The UDM relevance weights correspond to the probability that at least either a random assessor, or the observed one, would consider a particular result a top result, given the relevance level assigned by the latter.[2] As a result, the weight assigned to a result assessed as top relevant becomes one, and the weight for levels assessed below the top level ($i < T$) corresponds to the probability $p_{T|i}$. The sum of the UDM relevance weights over a set of results is the expected number of results based on the UDM user model. This corresponds to the expected number of results with at least one top level score by the observed assessor or a random one, and is obtained by adding up the actual number of results assessed with the top level, with fractional counts $p_{T|i<T}$ for lower rated results.

[2] The UDM was actually defined based on the probability that at least M out of N assessors, including the observed one, assign the top level. However, based on the binomial distribution, this is a straightforward extension from the case of $M = 1$ and $N = 2$, which is described here and corresponds best to the PRM formulation.

3.4.2 The PRM user model

The parameters of the PRM are based on a simpler and more intuitive user model, whereby we predict the relevance for a random user, again based only on the knowledge of an assessment level. By modeling a random user and leaving out the assessor, instead of selectively accepting those assessments with the highest level as true in the UDM, we do not enforce the top level relevance weight to be one, as in the UDM. The simplicity of the PRM user model leads to the interpretation of the summed disagreement parameters over a set of results as the *expected* number of relevant results for a random user.

3.4.3 Counter-intuitive results with the UDM

Although sound by itself, the original UDM user model is less intuitive than the new PRM user model, and may lead to counter-intuitive results in special situations. For example, consider the case where the top relevance level (T) and the second highest relevance level ($T-1$) are conceptually very close to one another (e.g., T defined as 'Top result', and $T-1$ as 'Excellent match', such that the distinction between both levels becomes really difficult for assessors). The confusion between these levels would yield both $p_{T|T-1} \approx 0.5$, and $p_{T|T} \approx 0.5$. Intuitively, both relevance levels could be considered top levels, and should therefore have similar weights. While the UDM assigns the weight for the official top relevance level T as 1, and approximately 0.5 for the other effective top level, the PRM would assign equal weights to both levels, following the intuition outlined above.

3.4.4 Link between user and assessor model

A further difference between the PRM and the UDM, is the link between the user and assessor model. Although the distinction between both models was made less explicit by Demeester et al. (2014) than in the current work, the UDM assumes that users are only satisfied with top results. The PRM is formulated in a more general way: relevance between users is defined separately from the assessment levels. As explained, it is convenient in practice if the various assessment levels can be mapped to the binary notion of user relevance. To this end, multiple choices are possible, with various interpretations of the evaluation scenario.

3.4.5 Gains for non-relevant results

In the UDM, the gain for the lowest relevance level ($i = 0$) was defined as zero, whereas the PRM gain of non-relevant results is the possibly non-zero value of $p_{R|0}$. Stating that results considered non-relevant by the assessor should have no contribution to evaluation metrics, as in the UDM, is convenient and in line with traditional evaluation strategies. However, we do not want to exclude situations where a random user might consider such a result relevant. In the PRM case, user relevance is not limited to the top assessment level as in the UDM. For example, if user relevance is captured by any assessment level above the lowest level (i.e., with threshold $\theta = 1$), confusion between user relevance R and the lowest assessment level becomes more likely, and can no longer be ignored in general.

In some cases, however, ignoring the contribution of $p_{R|0}$ is allowed, which allows significantly reducing the additional annotation effort for estimating the disagreement parameters (see Sect. 5.2.1). Also, when the disagreement parameters are meant to

represent the whole collection, e.g., obtained by randomly selecting documents for annotation, it may be convenient to explicitly set $p_{R|0}$ to zero, as in the UDM. The large majority of documents are most likely completely non-relevant to a given query, and should therefore not contribute to the total relevance. Low yet non-zero values of $p_{R|0}$ in this setting may be due to annotation errors. In practical scenarios, however, the disagreement parameters would often be estimated from a biased subset of the data, intended for evaluation purposes, e.g., by pooling search results. In most such cases, non-zero values of $p_{R|0}$ cannot be neglected. More details on how the disagreement parameters depend on the data subset used for evaluation, are given in Sect. 6.1.

4 Datasets

Before venturing into a more detailed analysis and discussion of the practical application of the PRM, we present the two data sets that we use to support that discussion with quantitative experiments, highlighting the properties and behavior of the PRM. Both datasets contain a (sub)set of double graded relevance assessments, and are as such ideal for experiments with the PRM.

4.1 TREC 2013: Federated Web Search Track

The first dataset used in this work comes from the TREC 2013 Federated Web Search Track (FedWeb13) (Demeester et al. 2013). This track was created to stimulate research in federated search and the dataset contains the actual results of 157 real web search engines, including both the returned snippets and the actual pages of the top-10 results for each query. The 2013 edition of the track featured a resource selection and results merging task. The goal of the resource selection task was to rank the different resources on their predicted relevance to the test topics. In the results merging task, participants had to create a single ranked list over the results from all resources. Although initially a large set of 200 test topics was provided to the participants, the evaluation itself was based on the judgements for 50 test topics.

Students with different backgrounds were recruited to judge the relevance of the results, covering the fields of engineering, law, computer science, music, economics, and arts. From the initial set of test topics, the students were assigned topics of their choice, according to their expertise, which they then had to entirely annotate. Although the queries were not judged by those who initially created the queries, they themselves wrote narratives on which the judgments were based, from their own perspective. Because they selected their own queries and defined the information need, it is reasonable to see them as primary rather than secondary assessors (see Sect. 3.1).

The relevance of search results was graded on the following levels: Non (not relevant), Rel (minimal relevance), HRel (highly relevant), Key (top relevance), and Nav (navigational). For our experiments, we merged the few Nav labels into the Key category (this was also done for the official task evaluation, as the test topics were not navigational in nature). The dataset contains 34,010 results for the 50 test topics, for which both the page and the snippet were (independently) judged. In addition, a subset of double judgments was collected for a subset of the data (6253 for the snippets and 7027 for the pages). These double judgments were mostly chosen at random, also depending on the availability of assessors. Sometimes only a few (e.g., the first three) results from a result list were judged

twice, sometimes all 10. In total, 26 of the test topics contain double snippet judgments, and 24 topics have double page judgments. The assessors that provided the second set of judgments for a particular query did not create the query or narratives themselves, but again judged queries they themselves could have created, according to their interests. As a result, the user population towards which the PRM parameters will be tuned, consists of students whose information needs are mostly informational.

Further information on the data and the relevance judgments can be found in the FedWeb13 overview paper (Demeester et al. 2013). All judgments were released by the track organizers in the 'Fedweb Greatest Hits Collection' (Demeester et al. 2015).

For the FedWeb13 system evaluation experiments described in Sect. 7, the 18 submitted runs by 9 teams for the resource selection task are used, as well as the 15 runs by 6 teams for the results merging task. The evaluation measures are calculated with the trec-eval software.[3]

4.2 NTCIR-10 2013: Intent-2 task

The second dataset used in this paper contains the relevance judgments for the NTCIR-10 INTENT-2 Task, more in particular the Document Ranking Subtask for Chinese and Japanese data. In this task, the participants were asked to return a ranked list of search results. The test queries were in part navigational in nature, and in part informational. For the latter, the participants were required to diversify their results to cover different navigational intents. The results to be manually judged were selected by means of fing over the submitted runs, with a pool depth of 40. For these, full double judgments are available, both for the Chinese test topics (22 navigational, and 75 informational ones), and for the Japanese topics (of which 28 are navigational and 67 informational). All judgments were done on a three-level scale, with levels 0 (non-relevant), 1 (medium), and 2 (highly relevant), and by hired assessors. For the evaluation, these paired judgments were combined into a set of single 5-level gains. The resulting reference set of 5-level labels contains 22,552 explicit Chinese judgments, and 13,172 Japanese ones. In the current paper, we only consider the double three-level judgments. More details can be found in the INTENT-2 overview paper by Sakai et al. (2013), and the overview paper at the first INTENT task at NTCIR-9, by Song et al. (2011), which provides additional details.

For the INTENT-2 evaluation experiments, the 12 submitted Chinese runs from 3 teams and the 8 submitted Japanese runs from 2 teams for the document reranking subtask are used, in combination with the NTCIREVAL toolkit.[4]

5 Practical calculation of the relevance gains

The following section describes a standard IR evaluation scenario for which the PRM method applies. A detailed description of how the disagreement parameters can be calculated, is given in Sect. 5.2, first in general, then in practice for the FedWeb13 and INTENT-2 data.

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

[4] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

5.1 General recipe

Compared to an evaluation scenario with a common ad-hoc choice of relevance level gains, the PRM comes with an extra annotation cost: it relies on additional judgments on a subset of search results, which are necessary to estimate the degree of assessor disagreement. The steps of the PRM approach are:

1. Gather a single set of graded relevance judgments for the test topics.
2. Optionally: if the test queries can be naturally divided into homogeneous subsets (such as informational and navigational queries), the disagreement can be separately modeled for them, and the evaluation setup separated. To this end, perform Step 3 for each of these subsets individually.
3. Perform the following steps on the data:
   (a) Gather a second set of judgments for a subset of the previously annotated search results, each by another assessor than for the original judgment.
   (b) Estimate the disagreement parameters $p_{T|i}$ for all relevance levels i (see Sect. 5.2).
   (c) Apply these as gains in suitable evaluation metrics (see Sect. 3.3).

The possibility mentioned in step 2 of dividing the data into more homogeneous subsets (e.g., according to different types of queries) has the advantage that a possibly different disagreement behavior is more accurately reflected in the different sets of relevance weights, as will be illustrated in Sect. 5.2. It however requires sufficient double annotations for each of these subsets, which makes it more costly in return.

Another important point pertains to the selection of a subset of search results to be annotated a second time in step 3a, from which the parameters $p_{T|i}$ will be determined. The distribution of the results (in terms of general search result quality) and the required number of double judgments are discussed in Sects. 6.1 and 6.3, respectively.

5.2 Estimation of the disagreement parameters $p_{T|i}$

5.2.1 General strategy

The discussion below covers the case where manual judgments are expensive, and at most two judgments from different assessors can be gathered for a subset of the test results. For the case of crowd-sourced Web search judgments, Demeester et al. (2014) show that the case of multiple judgments per result is approximated quite well by using only double judgments to estimate $p_{T|i}$. If more than two judgments per result are available, the formulas proposed below to estimate $p_{T|i}$ as a ratio of occurrence frequencies can be extended.

The two different sets of annotations for the chosen results subset (see Sect. 5.1), are denoted as the set from user (or user group) U_1 and the one from user (group) U_2. U_1 and U_2 may represent actual groups of assessors, or correspond to an arbitrary separation of each double judgment into two groups, if the double judgments were provided by a single group of assessors.

As explained in Sect. 3.1, the PRM relies on the assessors' capability of estimating user relevance. In practice, the assessment levels are defined such that the binary notion of user relevance can be obtained directly from them, e.g., based on a threshold θ. To keep the notations simple, we write 'relevant according to the binary user model' as R, denoting

Table 1 Illustration of using Eqs. (4) and (5) for estimating disagreement parameters based on 20 pairs of 3-level judgments (with labels 0, 1, or 2) by assessors U_1 and U_2

Results	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}	d_{17}	d_{18}	d_{19}	d_{20}
Ass. U_1	2	2	1	0	1	1	0	1	2	1	1	1	2	0	1	0	1	1	0	0
Ass. U_2	1	2	1	1	2	0	0	1	2	2	0	2	1	0	1	2	0	1	0	0
Eq. (4)						Eq. (4)							Eq. (5)							
$p_{U_2=2\|U_1=2} = 2/4$						$p_{U_1=2\|U_2=2} = 2/6$							$p_{2\|2} = 4/10$							
$p_{U_2=2\|U_1=1} = 3/10$						$p_{U_1=2\|U_2=1} = 2/7$							$p_{2\|1} = 5/17$							
$p_{U_2=2\|U_1=0} = 1/6$						$p_{U_1=2\|U_2=0} = 0/7$							$p_{2\|0} = 1/13$							

either of the levels on or above the threshold, or $i \geq \theta$. Over the double judgments on all considered test topics, we define $N_{U_2=R,U_1=i}$ as the number of results judged above the binary relevance threshold by U_2 and with level i by U_1, and $N_{U_1=i}$ as the number of results judged with assessment level i by U_1. We can estimate $p_{R|i}$ as

$$p_{R|i} = \frac{N_{U_2=R,U_1=i}}{N_{U_1=i}}. \tag{4}$$

If both groups of assessors independently judged the same pool of results, an alternative estimation is given by

$$p_{R|i} = \frac{N_{U_1=R,U_2=i} + N_{U_2=R,U_1=i}}{N_{U_2=i} + N_{U_1=i}}. \tag{5}$$

In order to make the estimation procedure more tangible, Table 1 illustrates the use of Eqs. (4) and (5) for estimating disagreement parameters. In an artificial setting with 20 documents (d_1 to d_{20}), the 3-level graded relevance judgments by assessors U_1 and U_2 with respect to a query are listed, followed by the different estimates of disagreement parameters $p_{2|i}$ with respect to the top level 2. Note that in reality the counts need to be higher, in order to obtain good estimates.

The PRM is based on the assumption that the average disagreement only depends on the observed relevance level, and is independent of the particular assessor. In reality, for the latter Carterette and Soboroff (2010) have shown that assessors may actually differ in the proportion of documents they find relevant. For two such users U_1 and U_2, that would lead to a difference between $p_{U_1=R|U_2=i}$ and $p_{U_2=R|U_1=i}$, while Eq. (5) takes into account a higher number of judgments and leads to an averaged estimate.

If U_1 and U_2 each contain judgments from multiple assessors, Eq. (5) is still more robust, but using Eq. (4) would be sufficient. In some situations, the amount of required double judgments can be strongly reduced. If during the second assessment round it becomes apparent that the estimate of parameter $p_{R|0}$ is negligible, the extra judgments (by U_2) can be continued on a subset of only those indicated above non-relevance by U_1. This is illustrated in Sect. 5.2.2. Note that in this case, Eq. (5) is no longer valid and the one-sided estimation Eq. (4) must be used, because the distribution of the relevance levels by U_2 no longer corresponds with the one from U_1. For example, a large fraction of level 1 judgments by U_2 would be missing (correlated with those indicated with level 0 by U_1), whereas most top level judgments would be present, such that the estimate in Eq. (5) would be artificially high.

Table 2 Estimated $p_{R|i}$ ($\pm$ 1 std.) for top relevance ($\theta = \mathsf{Key}$) on the FedWeb13 data, for pages and snippets, and using Eq. (5) versus (4)

FedWeb13	Pages (5)	Pages (4)	Snippets (5)	Snippets (4)
$p_{\mathsf{Key}\|\mathsf{Key}}$	0.53 $\pm$ 0.02	0.52 $\pm$ 0.02	0.47 $\pm$ 0.02	0.49 $\pm$ 0.03
$p_{\mathsf{Key}\|\mathsf{HRel}}$	0.27 $\pm$ 0.01	0.28 $\pm$ 0.02	0.25 $\pm$ 0.01	0.26 $\pm$ 0.02
$p_{\mathsf{Key}\|\mathsf{Rel}}$	0.04 $\pm$ 0.01	0.05 $\pm$ 0.01	0.08 $\pm$ 0.01	0.09 $\pm$ 0.01
$p_{\mathsf{Key}\|\mathsf{Non}}$	0.01 $\pm$ 0.00	–	0.01 $\pm$ 0.00	–

Alternative estimates for $p_{R|i}$ can be devised, e.g., with smoothing based on a Dirichlet prior, to deal with low numbers of occurrence of certain label combinations. This is left open for future research.

5.2.2 FedWeb13 disagreement parameters

Table 2 shows the estimated disagreement parameters for the FedWeb13 data, both for pages and snippets, for the case that users are only satisfied with top results ($R = \mathsf{Key}$). There appears to be a substantial confusion between both highest levels (Key and HRel), and less so for the lower levels. The disagreement on different levels is similar for the full pages and for the snippets. The standard error on these estimates is shown as well, which is highest for $p_{\mathsf{Key}|\mathsf{Key}}$ (as the combination of two Key judgments occurs the least), but remains within a few per cent.

Note that the standard deviation σ on the estimate of $p_{R|i}$ can be estimated as follows, given that $p_{R|i}$ is the success rate in a binomial distribution:

$$\sigma_{p_{R|i}} = \sqrt{\frac{N_N}{N_D}\left(1 - \frac{N_N}{N_D}\right)\frac{1}{N_D}},$$

where N_N and N_D represent the numerator and denominator, respectively, of Eqs. (4) or (5), depending on the estimation method.

The estimates based on Eq. (5) are compared in Table 2 with those based on the one-sided estimate Eq. (4). For the latter, only rejudgments of results originally judged above Non were used, such that $p_{\mathsf{Key}|\mathsf{Non}}$ could not be estimated. Neglecting the contribution from the lowest level is allowed in this case, due to the very low confusion with respect to the top level ($p_{\mathsf{Key}|\mathsf{Non}} = 0.01$). The results calculated with Eq. (5) are slightly more robust (i.e., they take into account more top judgments, and display a lower standard deviation). However, the differences are small, and the total number of double page (snippet) judgments used for the one-sided estimation amounts to only 19 % for pages and 23 % for snippets compared to the estimates with Eq. (5).

Table 3 provides the disagreement parameters for the FedWeb13 page judgments, for three different choices of the threshold that defines binary user relevance as a function of the assessment levels. The left column shows results for binary relevance on the Key level ($\theta = \mathsf{Key}$), the middle column assumes that users are satisfied with results they think satisfy the descriptions of either Key or HRel (with threshold $\theta = \mathsf{HRel}$), and the right column assumes that all levels above Non are relevant to the user ($\theta = \mathsf{Rel}$). Relaxing the notion of user relevance leads to larger probabilities $p_{R|i}$. For example, where only 53 % of the users would consider a result assessed with the Key label effectively a key result, 93 % would consider it at least marginally relevant. For the recall-oriented user scenario $\theta = \mathsf{Rel}$, an observed assessment with label HRel is almost as likely as a Key assessment

Table 3 $p_{R|i}$ estimated from FedWeb13 page judgments, for different thresholds θ of binary user relevance R

FedWeb13	$\theta = \mathsf{Key}$	$\theta = \mathsf{HRel}$	$\theta = \mathsf{Rel}$
$p_{R\|\mathsf{Key}}$	0.53	0.87	0.93
$p_{R\|\mathsf{HRel}}$	0.27	0.65	0.88
$p_{R\|\mathsf{Rel}}$	0.04	0.22	0.46
$p_{R\|\mathsf{Non}}$	0.01	0.02	0.08

to lead to relevance for a random user. For the precision-oriented approach $\theta = \mathsf{Key}$, results assessed as HRel are only half as likely to satisfy a user as results assessed with the Key label. Also note that $p_{R|\mathsf{Non}}$ is very small for $\theta = \mathsf{Key}$, due to the limited confusion between the top and the lowest level, whereas it is larger for $\theta = \mathsf{Rel}$, due to the disagreement between the levels Non and Rel.

5.2.3 INTENT-2 disagreement parameters

Table 4 shows the estimated disagreement parameters $p_{2|\mathsf{i}}$ for the INTENT-2 data, estimated from all double 3-level judgments (column 'all' in the table). When we consider all queries together, there seems to be a fair agreement on the top level. The amount of confusion between the highest and the middle level (i.e., $p_{2|1}$) is however much higher for the Japanese than for the Chinese data. This disagreement on the Japanese queries was already noticed by Sakai et al. (2013), without giving any rationale behind it.

Two query types can be distinguished: navigational (22 Chinese and 28 Japanese queries) and informational (75 Chinese and 67 Japanese queries). The informational queries contribute more strongly to the combined results (column 'all') than the navigational ones, because there are more of them, and they have multiple intents. To investigate the influence of the navigational queries, we also calculated the disagreement parameters separately for the different query types, in agreement with step 2 of the general recipe (Sect. 5.1). Table 4 illustrates clearly that these query types lead to a very different disagreement in both languages, such that making this distinction is justified and necessary.

For the navigational queries (column 'nav.'), there is a large difference in the level of agreement on the top results between both languages, which is very high for Japanese, and very low for Chinese. For the latter, the fact that $p_{2|2}$ is so small, shows that there may be a problem with the relevance judgments, or at least with the assessors' interpretation of the top relevance level for a navigational query.

For the informational queries, where multiple intents of the same query were separately judged, we considered different approaches to estimate the disagreement parameters. For the first approach (indicated as 'all intents' in Table 4), we considered each given (query, intent) pair as a different information need. The double judgments over all different intents

Table 4 Estimated $p_{T|i}$ ($\pm$ 1 std.) on different query types (all, navigational, informational) for the INTENT-2 data

INTENT-2	All	Nav.	Inf. (all intents)	Inf. (top intent)
Japanese				
$p_{2\|2}$	0.51 $\pm$ 0.01	0.77 $\pm$ 0.02	0.49 $\pm$ 0.01	0.54 $\pm$ 0.01
$p_{2\|1}$	0.19 $\pm$ 0.00	0.17 $\pm$ 0.02	0.20 $\pm$ 0.01	0.24 $\pm$ 0.01
$p_{2\|0}$	0.03 $\pm$ 0.00	0.03 $\pm$ 0.00	0.03 $\pm$ 0.00	0.05 $\pm$ 0.00
Chinese				
$p_{2\|2}$	0.37 $\pm$ 0.01	0.07 $\pm$ 0.02	0.41 $\pm$ 0.02	0.29 $\pm$ 0.03
$p_{2\|1}$	0.03 $\pm$ 0.00	0.04 $\pm$ 0.00	0.04 $\pm$ 0.00	0.04 $\pm$ 0.00
$p_{2\|0}$	0.00 $\pm$ 0.00	0.01 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00

and queries were taken together, and the parameters $p_{2|2}$ were calculated with Eq. (5). Note that the judgments with intent label '0' (meaning none of the intents were judged relevant), were replaced by explicit separate judgments of non-relevance for each of the intents. The second approach ('top intent') is based on only the most probable intent for each query, given the intent probabilities (as in Sakai et al. 2013). The underlying idea is that the most probable intent for a query may lead to a different disagreement behavior than the average over all intents. Since even considering the top intents alone leads to enough judgments for confident estimates, Step 2 of the general recipe can be applied. For the Japanese data, the behavior remains the same, except for a small increase in the overall probability on a top judgment. For the Chinese data, there is an overall increase in the disagreement (lower $p_{2|2}$ and higher $p_{2|1}$). In the remainder of the paper, the disagreement parameters as estimated from the top intents will be used. The reason is that for the evaluation part (Sect. 7) the influence of the disagreement on the nDCG metric will be investigated, i.e., to evaluate results on a single intent, as opposed to more advanced variations that account for intent diversity.

6 Analysis of PRM parameters

In the following sections we take a closer look at some general properties and difficulties in applying the PRM, integrating experimental evidence immediately into the discussions. These issues include the dependence of the PRM parameters on the results quality (Sect. 6.1), the choice of test topics (Sect. 6.2), and the number of double judgments (Sect. 6.3).

6.1 Sensitivity to search result quality

An important issue that may influence the final estimates for $p_{R|i}$ is the choice of the initial set of double annotations. Webber et al. (2012) show that assessor disagreement on particular documents depends on the ranks at which these documents are retrieved by a set of retrieval systems, summarized into their 'metarank': they model how the disagreement changes as a function of that metarank. This effect was also observed by Demeester et al.

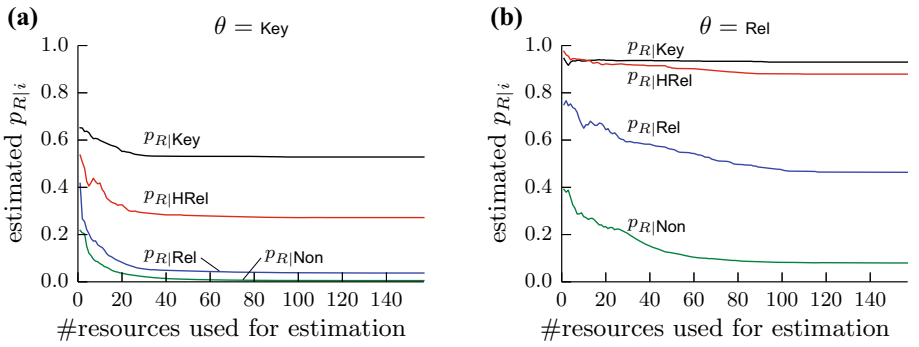

Fig. 1 Disagreement parameters for the FedWeb13 data, estimated using top-10 results from the top-k ranked resources (in decreasing order of results relevance). We used double annotations of all results considered for the estimation

(2014) for the FedWeb12 dataset: if the $p_{\mathsf{Key}|i}$ parameters are estimated from high-quality results lists, they are larger than when estimated from average result lists. We now further explore this effect.

Figure 1 visualizes the described phenomenon for the FedWeb13 data. For each query, we ordered all resources (i.e., search engines) according to the descending number of Key or HRel results they returned (as measured by the single reference assessor for which full judgments are available). This leads to an ad-hoc ranking from high-quality to low-quality resources. The $p_{R|i}$ curves in Fig. 1 were obtained by gradually taking into account the top-10 results from more resources, starting from only the best resource, up to including them all. Two different scenarios for user relevance are shown: (a) with threshold $\theta = \mathsf{Key}$ corresponding with users that are only satisfied with top results, and (b) for $\theta = \mathsf{Rel}$, for users that are satisfied with any result at least marginally relevant. The asymptotic values, when all resources are taken into account, correspond to the values listed in Table 3. Note that in Sect. 7 we will further use these user scenarios, referring to them as the top relevance scenario ($\theta = \mathsf{Key}$) and the marginal relevance scenario ($\theta = \mathsf{Rel}$).

The disagreement parameters start high, when only high-quality resources are used, then decrease, and finally saturate as soon as the lower ranked resources contain no further results with the appropriate relevance levels to contribute. We would like to stress the fact that the judgments were done in a randomized order (within each query), where the assessors were not informed on the provenance (i.e., resource) of the web page under assessment. This means that indeed the effect described by Webber et al. (2012) can be observed. One possible explanation for the observed behavior is due to the limitations of using a small discrete set of relevance levels. Consider for example the observed behavior of $p_{R|\mathsf{Key}}$, for the relevance threshold $\theta = \mathsf{Key}$. Among all results assessed as Key, we can imagine that some would receive an even higher relevance grade if it existed, with a correspondingly higher probability of an average user to consider it relevant. Such results considered more relevant than the average results judged as Key, are more likely to come from the best resources, hence the elevated levels of $p_{R|\mathsf{Key}}$ if only these are taken into account. For the case $\theta = \mathsf{Rel}$, this effect on $p_{R|\mathsf{Key}}$ is very small, because apparently the explained variations among results indicated as Key do not strongly influence the probability of a user to consider a result at least marginally relevant.

The take-away message of this discussion is the following. We have seen that the disagreement parameters may vary, depending on the set of results they are estimated from. Therefore, the main consideration for defining the set of double annotations to estimate $p_{R|i}$ from, is that it should be representative for the evaluation setting. For setups where the evaluation targets the higher-ranked results, those should be sampled from when gathering the double relevance judgments. For example, in the case of pool-based IR evaluation, if results up to a depth of 10 will be used for measuring system comparisons, a reasonable choice would be to gather the double annotations from a sample of the top-10 results by the systems under test.

For the experimental results shown in this paper, we have chosen to use the same disagreement parameters for the different evaluation settings, based on all available double judgments.

6.2 Choice of test topics

The disagreement parameters are calculated by aggregating double judgments over multiple test topics. However, the disagreement between assessors might depend on the

particular topics, and yet the same PRM parameters are used for all topics. In particular, for the FedWeb13 data, approximately one out of five results was judged twice, distributed among half of the test topics (see Sect. 4.1), and the relevance gains based on those are used to evaluate all 50 topics. In order to visualize the dependence on the topics, we bootstrapped the different topics for which double annotations were performed, each time estimating the disagreement parameters based on the selected topics, in 300 bootstrap samples. For the calculations, all judgments for each of the topics were taken into account as many times as the topic was chosen for the particular bootstrap sample. Figure 2 shows a boxplot of the result, both for snippets and pages, with a similar behavior. The largest variation occurs for the highest relevance levels, because their estimates are based on the fewest cases. For example for the pages, we find a standard deviation of 0.06 on the estimate of $p_{\mathsf{Key}|\mathsf{Key}}$ and 0.05 on $p_{\mathsf{Key}|\mathsf{HRel}}$. These values are higher than the corresponding standard deviations due to the total number of cases to estimate the disagreement parameters from, which are 0.02 and 0.01, respectively (see Table 2). We conclude that the influence of the topics is noticeable, but does not invalidate the disagreement parameters because the variation is still limited. However, for using the PRM method, we recommend to gather incomplete sets of double judgments for a larger fraction of the test topics, as was done for the FedWeb13 data, rather than complete double judgments on a smaller number of topics, as previously done for the FedWeb12 data (see Demeester et al. 2014).

6.3 Number of double judgments

We now discuss the required number of double judgments. Given their extra annotation cost, ideally the number of double judgments should be kept to a minimum. The main requirement is that there are enough judgments to have a small enough uncertainty on the disagreement estimates. The allowed upper boundary of that uncertainty depends on the application. Yet, requiring that that the disagreement parameters for levels with a conceptually clear difference in relevance are well distinguishable, can be used as a sufficient condition for the number of double judgments. Both for the FedWeb13 and INTENT-2 data, the standard deviations (shown in Tables 2, 4) are small enough in that respect. The only exception is the vague distinction between the top and medium level for the Chinese navigational queries, due the very low top level agreement.

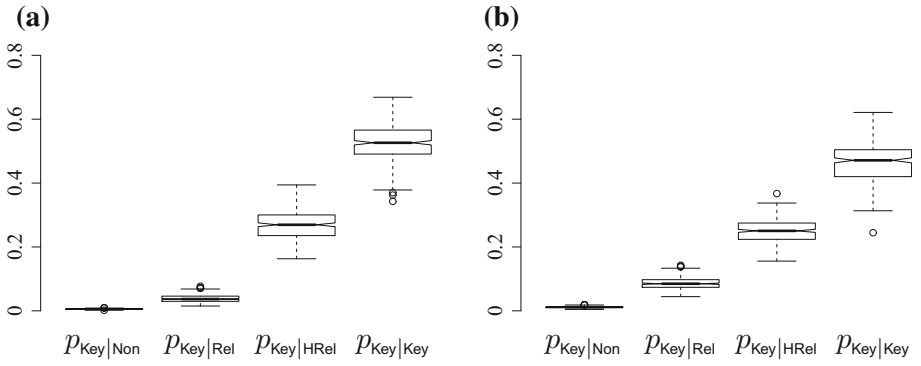

Fig. 2 Boxplot of $p_{T|i}$ for different levels i, by bootstrapping the test topics for the FedWeb13 pages (**a**) and snippets (**b**)

To get an idea of the uncertainty on the estimates as a function of the required number of double judgments, we did the following bootstrap experiment on the INTENT-2 data, focusing on the scenario with paired judgments and Eq. (5) to estimate $p_{2|i}$. We simulated 50 annotation rounds by sampling the actual double judgments (with replacement), keeping track of the disagreement parameters for the growing set of simulated double judgments. The mean value and one standard deviation above and below it are shown in Fig. 3. Given the large amount of judgments (i.e., full double judgments), the uncertainty on most of the estimates already becomes very small for a fraction of the judgments. For the Chinese navigational queries, however, the problems noted in Sect. 5.2.3 are confirmed. Even when the absolute uncertainty on $p_{2|1}$ and $p_{2|0}$ becomes small, they cannot be distinguished in terms of their disagreement with the top level. This makes the resulting parameters $p_{2|i}$ less trustworthy, and any evaluation based only on these queries questionable.

7 Application of the PRM for system evaluations

This section is devoted to the application of the PRM model to actual system evaluations, based on the INTENT-2 and FedWeb13 data. We will demonstrate the difference in counting the number of relevant results purely based on the assessor and as expected for a random user (Sect. 7.1), demonstrate the robustness of evaluation with the PRM (Sect. 7.2), and investigate the behavior of system rankings based on PRM gains versus heuristic gains (Sect. 7.3).

7.1 Counting relevant results

As explained in Sect. 3.2, summing the disagreement parameters $p_{R|i}$ for each result in a result list, according to the assigned relevance level by the assessor, results in the expected number of relevant results according to a random user. This allows making absolute

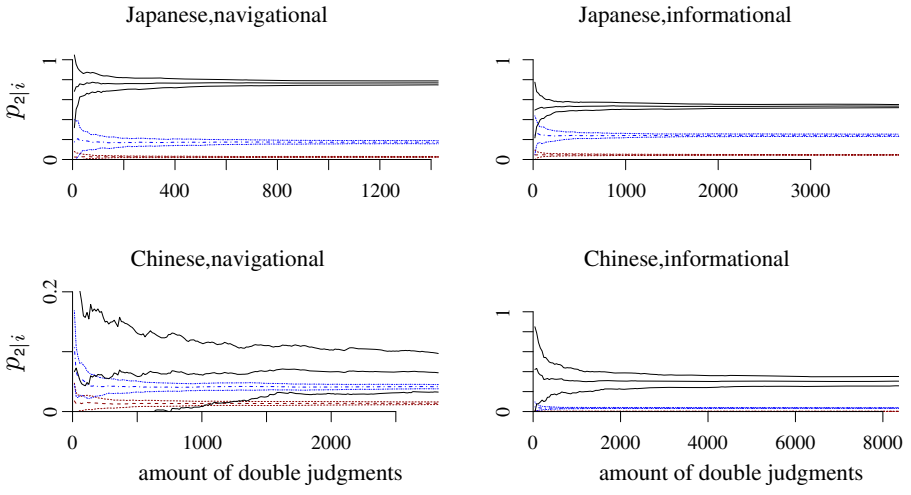

Fig. 3 Simulated $p_{2|i}$ (mean $\pm$ 1 std.) on INTENT-2 judgments versus the number of double judgments: $p_{2|2}$ (*black full lines*), $p_{2|1}$ (*blue dash-dot lines*), $p_{2|0}$ (*red dashed lines*) (Color figure online)

conclusions about how well systems are able to return relevant results, whereas the ad hoc weighting (e.g., linear or exponential) of relevance levels typically focuses on relative system comparisons, without a clear interpretation of the absolute value of the resulting metrics.

Let us illustrate this with the results of the FedWeb13 Resource Selection (RS) task. Participants were required to rank 157 online resources on their estimated capability of returning relevant results for a particular query. From 9 participating teams, results for 18 RS systems were submitted. In a typical federated search scenario, the results from the highest ranked resources per query are retrieved, merged into a single ranked list, and presented to the user. We only consider the top three resources per query. Given that per query, only the top-10 search results for each resource are available, the result sets that we evaluate for each system contain at most 30 results per query.

Table 5 shows the number of relevant results among the top three resources, averaged over 50 evaluation queries, together with the standard deviation on that average. The columns 'PRM' show the number of relevant results a random user expects to find, estimated with Eq. (2) according to the PRM model. The columns 'binary' show the number of

Table 5 Number of relevant results among top 3 resources, for FedWeb13 Resource Selection runs (average over 50 test queries ± the st. dev. of the mean). PRM: expected number of relevant results for a random user; binary: number of relevant results by a single assessor. User scenarios: top relevance versus marginal relevance

run	Top relevance ($\theta = \mathsf{Key}$)		Marginal relevance ($\theta = \mathsf{Rel}$)	
	PRM	Binary	PRM	Binary
oracle	9.02 (±0.40)	12.98 (±0.93)	21.40 (±0.61)	25.92 (±0.63)
RS_clueweb	2.51 (±0.27)	2.66 (±0.46)	8.52 (±0.54)	9.26 (±0.74)
UiSSP	2.41 (±0.36)	2.58 (±0.57)	7.78 (±0.82)	8.76 (±1.20)
UiSP	2.27 (±0.38)	2.52 (±0.63)	7.22 (±0.84)	8.00 (±1.17)
utTailyNormM400	2.05 (±0.37)	2.20 (±0.54)	6.65 (±0.86)	7.24 (±1.14)
utTailyM400	1.94 (±0.37)	2.06 (±0.54)	6.32 (±0.86)	6.74 (±1.13)
UiSS	1.66 (±0.25)	1.64 (±0.35)	5.84 (±0.65)	5.98 (±0.92)
udelODRA	1.63 (±0.30)	1.68 (±0.48)	5.64 (±0.73)	6.00 (±1.01)
udelFAVE	1.63 (±0.29)	1.62 (±0.43)	5.78 (±0.73)	6.14 (±1.01)
UPDFW13mu	1.54 (±0.32)	1.52 (±0.44)	5.18 (±0.79)	5.36 (±1.06)
iiitnaive01	1.46 (±0.28)	1.52 (±0.45)	5.24 (±0.65)	5.76 (±0.88)
cwi13SniTI	1.46 (±0.29)	1.46 (±0.44)	5.17 (±0.70)	5.48 (±0.96)
UPDFW13sh	1.43 (±0.27)	1.12 (±0.32)	5.28 (±0.71)	5.38 (±0.98)
RS_querypools	1.25 (±0.19)	1.06 (±0.34)	5.33 (±0.44)	6.18 (±0.80)
cwi13ODPTI	1.17 (±0.21)	0.92 (±0.27)	4.48 (±0.56)	4.52 (±0.77)
ECNUBM25	0.91 (±0.22)	1.16 (±0.38)	3.04 (±0.64)	2.62 (±0.56)
cwi13ODPJac	0.66 (±0.16)	0.42 (±0.16)	2.88 (±0.51)	2.94 (±0.72)
udelRSMIN	0.61 (±0.21)	0.78 (±0.33)	2.28 (±0.48)	1.94 (±0.62)
incgqdv2	0.55 (±0.13)	0.48 (±0.21)	2.17 (±0.35)	2.08 (±0.46)
incgqd	0.35 (±0.12)	0.30 (±0.19)	1.46 (±0.29)	1.38 (±0.40)
StanfordEIG10	0.19 (±0.07)	0.14 (±0.08)	0.85 (±0.20)	0.66 (±0.27)

relevant results based on Eq. (1), purely based on single judgments, i.e., ignoring the disagreement. Two different scenarios for user relevance are shown, the top relevance scenario ($\theta = \mathsf{Key}$), and the marginal relevance scenario ($\theta = \mathsf{Rel}$).

The results are shown for the 18 official runs,[5] as well as two baselines by the organizers (`RS_clueweb` and `RS_querypools`), and an artificial RS system (`oracle`) that selects the three best possible resources per query.

We can make a number of observations from these results. For both user scenarios, the PRM estimates of the average number of relevant results are more robust, given the lower standard errors, than the binary estimates. The system rankings between PRM and binary estimates are strongly correlated, although not the same: Kendall's tau is 0.93 for the top relevance scenario, and 0.89 for the marginal relevance scenario.

We observe substantial differences in the absolute numbers of estimated relevant results, due to the difference between modeling disagreement (PRM) and accepting the assessors' judgments as ground truth (binary). Based on the disagreement parameters, these differences can be interpreted. For example in the user scenario of top relevance, two main effects play a role in the PRM results: (1) The strong disagreement on the top level ($p_{\mathsf{Key}|\mathsf{Key}} = 0.53$) causes results assessed as Key to contribute only half as much to the estimated number of top relevant results, compared to the binary estimate; (2) Results only assessed as HRel are considered Key results by random users in about one out of four times ($p_{\mathsf{Key}|\mathsf{HRel}} = 0.27$). For the oracle system, the top 3 resources contain 13 Key results, purely based on the assessor, whereas a random user expects to find only 9 Key results. This means effect (1) is dominant. Some of the lower ranked systems have a higher PRM-based than binary estimate of the number of Key results, for example the run `cwi130DPJac`. For such systems effect (2) dominates, and they are better at retrieving results assessed as HRel than Key results.

7.2 Robustness of PRM-based evaluation

A direct way to evaluate how well a system is capable of retrieving relevant documents, is by calculating effectiveness measures based on binary relevance: relevant results are rewarded, depending on the rank at which they are retrieved. Due to user disagreement on the top level, the evaluation scores and even score-based rankings between different systems may lack robustness. The PRM allows us to reward results based not only on the particular assessor's personal idea of user relevance, but on the expected relevance to a random user. Because the latter is estimated from the average disagreement between assessors, a PRM-based evaluation should lead to a more robust evaluation, with respect to the choice of assessors.

This can be verified with the double set of 3-level INTENT-2 judgments, and the official runs submitted to the INTENT-2 Document Ranking Subtask. We consider user relevance at the highest assessment level $\theta = 2$: our evaluation reflects users that are only satisfied with top results. Each run is scored separately for the set of judgments from user U_1 and from user U_2. As an indicator of robustness, we consider Kendall's rank correlation coefficient τ between the resulting rankings of the runs, each based on one of the sets of assessments, i.e., U_1 versus U_2. As evaluation measure, we use nDCG@10, averaged over the test topics, and with a logarithmic discount function. Table 6 lists the results for the binary nDCG as introduced in Sect. 3.3.2 (column 'binary'), for the PRM-based nDCG in

[5] The TREC results are available at http://trec.nist.gov/results/.

Table 6 Kendall τ between system rankings based on different users for the INTENT-2 data, based on nDCG@10 with binary gains on top relevance, corresponding PRM gains, and linear gains

INTENT-2	Binary	PRM	Linear
Japanese nav.	0.86	0.86	0.64
Japanese inf.	0.84	0.93	1.00
Chinese nav.	0.12	0.43	0.47
Chinese inf.	0.06	0.27	0.72

which the disagreement parameters $p_{2|i}$ are used as gains (column 'PRM'), and with linear gains $g(i) = i$ (column 'linear').

There is a clear difference between the Chinese and Japanese data: the order of the Japanese runs seems almost user-independent, whereas there is a strong mismatch for the Chinese data. This may in part be related to the limited amount of data: only 8 Japanese runs from 2 teams, and 12 Chinese from 3 teams. Another cause may be the organization of the assessments, since U_1 and U_2 actually contain judgments from multiple judges, but we cannot further investigate this effect, as the composition of U_1 and U_2 for both languages has not been made public. Yet, the main reason is the higher overlap on top judgments for the Japanese data, as opposed to the Chinese: we have $p_{2|2} = 0.77$ and 0.54 for respectively the navigational and the informational queries in the Japanese data, while the Chinese has only $p_{2|2} = 0.07$, respectively 0.29.

The robustness of the evaluation based on U_1 and U_2 is finally also tested with linear gain values. In this case, the robustness also increases significantly with respect to the binary top evaluation. However, it is important to stress that there is an important conceptual difference between using the PRM and using linear gains. The choice of linear gains may be defensible in certain scenarios, but does not allow specifically testing the capabilities of a system in retrieving top relevant results, which both the top binary evaluation scenario and the associated PRM scenario do. For example, for the Chinese informational queries the linear gains lead to a higher robustness than the PRM, due to the stronger weighting of the medium levels. In this case the linear gain for the medium relevance level equals half the gain of the highly relevant results. What does this mean for the evaluation scenario? The PRM gain of level 1, i.e., the chance that a random user would assign 2 if the assessor said 1, is actually much lower than half the top level gain, or the corresponding chance if the assessor had said 2: a fraction 0.12. In other words, the linear gain of the medium level is too high to only account for disagreement on the top level. In this example, evaluation with linear gains not only rewards systems for retrieving top results, it also rewards them for their capability in retrieving medium results. Evaluation with linear gains is therefore not in line with the user model behind the binary and PRM gains, i.e., user relevance for top results.

A disadvantage of using fixed heuristic gains, is that interpretations as the one above are data-dependent. For example, in situations with very high disagreement, a linear gain might even not be high enough to compensate for the confusion of a particular level with the levels $i \geq \theta$. The PRM, in contrast, has an underlying evaluation scenario with a direct interpretation.

Table 7 FedWeb13 Results Merging evaluation: Kendall τ between nDCG@20 based system rankings for different sets of gains, and two user relevance scenarios: top relevance versus marginal relevance

FedWeb13	Top relevance	Marginal relevance
PRM versus binary	0.75	0.94
PRM versus linear	0.96	0.92
PRM versus exponential	0.99	0.93

7.3 Evaluation with PRM gains versus standard gains

We now consider the TREC FedWeb13 Results Merging Task, in which participants were challenged to design algorithms to create a merged ranking of the top-10 results from 157 online search engines. The official metric was nDCG@20, and for the evaluation, only the first of any returned duplicates was taken into account. We used the same evaluation methods, but altered the gains used for nDCG. Table 7 shows Kendall's τ between rankings of the official results merging runs based on different sets of nDCG gains: binary, PRM, linear, and exponential. For the top relevance scenario, the difference between the PRM and the binary relevance indicates the necessity of compensating for disagreement ($\tau = 0.75$). However, using the PRM, linear, or exponential gains seems to make little difference. In the marginal relevance case, the influence of using the PRM versus binary weights is much smaller ($\tau = 0.94$). The PRM-based ranking is still highly correlated with the rankings based on exponential or linear gains, although less than in the top relevance scenario. This is due to the stronger influence of the lower level PRM gains in the marginal relevance scenario.

8 Conclusions and future work

In this paper, we presented and analyzed the Predicted Relevance Model (PRM), which allows evaluating relevance towards a random user instead of purely accepting assessments as ground truth. The PRM allows quantifying the relevance for a random user, associated with multiple graded or categorical assessment levels, based on the disagreement between assessors. It was shown how existing evaluation measures can benefit from the PRM, leading to a robust evaluation of search engines with respect to several possible notions of binary user relevance, linked with the assessment levels. In a series of experiments based on existing evaluation collections, we explained how the PRM can be applied in practice, and analyzed its properties in actual evaluation scenarios.

This paper opens up several possibilities for future research. One straightforward direction is in further studying how the PRM can be applied to graded relevance evaluation measures other than the nDCG, or in other scenarios of user relevance. Another logical next step is the development of a principled way to combine the original view of graded relevance judgments as a measure of fractional utility, with the PRM ideas based on disagreement probabilities and binary user relevance. Furthermore, the PRM covers only one particular aspect of the general pursuit of predicting relevance of results towards users, namely the influence of disagreement. Other aspects that could be taken into account are, for example, the impact of multiple observed judgments per result, characteristics of

individual assessors or users, the type of test topics, the result snippet observed by the users, etc. The relevance of a result given a query, prior to observing one or more assessments, could for example depend on the type of query and the snippet shown to the users. Instead of the disagreement parameters according to the PRM, a more accurate posterior probability of relevance could be calculated after observing the available judgments on that particular result. We hope the insights gained in our current work will help in making progress towards this goal.

Acknowledgments First of all, we would like to thank the reviewers. Their particularly detailed comments and suggestions lifted the paper's overall quality and coherence, and played an important role in shaping the formulation and interpretation of the PRM in its current form. We would also like to thank Dolf Trieschnigg for his work on the FedWeb13 data and the many technical discussions, Tetsuya Sakai for providing us with the NTCIR INTENT-2 data, and Bart Deygers for the valuable suggestions to improve the manuscript. This work was supported by Ghent University—iMinds in Belgium, and by the Dutch national program COMMIT and the NWO-Catch project Folktales As Classifiable Texts (FACT) in the Netherlands.

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM international conference on web search and data mining* (*WSDM 2009*) (pp. 5–14), Barcelona. doi:10.1145/1498759.1498766.

Al-Harbi, A. L., & Smucker, M. D. (2014). A qualitative exploration of secondary assessor relevance judging behavior categories and subject descriptors. In *Proceedings of the 5th information interaction in context symposium* (*IIiX 2014*) (pp. 195–204), Regensburg. doi:10.1145/2637002.2637025.

Bailey, P., Craswell, N., Soboroff, I., & Thomas, P. (2008). Relevance assessment: Are judges exchangeable and does it matter? In *Proceedings of the 31st international ACM SIGIR conference research and development in information retrieval* (*SIGIR 2008*), Singapore. doi:10.1145/1390334.1390447.

Carterette, B., & Soboroff, I. (2010). The effect of assessor errors on IR system evaluation. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2010*) (pp. 539–546), Geneva. doi:10.1145/1835449.1835540.

Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. (2008). Here or there: Preference judgments for relevance. In *Proceedngs of the 30th European conference on advances in information retrieval* (*ECIR 2008*) (pp. 16–27). Berlin: Springer.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Incorporating variability in user behavior into systems based evaluation. In *Proceedings of the 21st ACM international conference on information and knowledge management (CIKM'12)* (pp. 135–144). New York, NY: ACM. doi:10.1145/2396761.2396782.

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management* (*CIKM 2009*) (pp. 621–630), New York, NY. doi:10.1145/1645953.1646033.

Demeester, T., Trieschnigg, D., Nguyen, D., & Hiemstra, D. (2013). Overview of the trec 2013 federated web search track. In *Proceedings of the 22nd text retrieval conference* (*TREC 2013*), Gaithersburg, MD.

Demeester, T., Aly, R., Hiemstra, D., Nguyen, D., Trieschnigg, D., & Develder, C. (2014). Exploiting user disagreement for web search evaluation: An experimental approach. In *Proceedings of the 7th ACM international conference on web search and data mining* (*WSDM 2014*) (pp. 33–42), New York, NY. doi:10.1145/2556195.2556268.

Demeester, T., Trieschnigg, D., Zhou, K., Nguyen, D., & Hiemstra, D. (2015). FedWeb greatest hits: Presenting the new test collection for federated web search. In *Proceedings of the 24th international world wide web conference* (*WWW 2015*), Florence. doi:10.1145/2740908.2742755.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, *47*(1), 37–49. doi:10.1002/(SICI)1097-4571(199601)47:1<3.0.CO;2-3.

Hosseini, M., Cox, I. J., Milić-frayling, N., Kazai, G., & Vinay, V. (2012). On aggregating labels from multiple crowd workers to infer relevance of documents. In *Proceedings of the 34th European conference on advances in information retrieval* (*ECIR 2012*) (pp. 182–194), Barcelona.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. doi:10.1145/582415.582418.

Kanoulas, E., & Aslam, J. A. (2009). Empirical justification of the gain and discount function for nDCG. In *Proceedings of the 18th ACM international conference on information and knowledge management* (*CIKM 2009*) (pp. 611–620), Hong Kong. doi:10.1145/1645953.1646032.

Kazai, G., Yilmaz, E., Craswell, N., & Tahaghoghi, S. (2013). User intent and assessor disagreement in web search evaluation. In *Proceedings of the 22nd ACM international conference on conference on information and knowledge management* (*CIKM 2013*) (pp. 699–708). New York, NY: ACM. doi:10.1145/2505515.2505716.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations: Comparison of the effects on ranking of IR systems. *Information Processing & Management*, *41*(5), 1019–1033. doi:10.1016/j.ipm.2005.01.004.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, *27*(1), 2:1–2:27. doi:10.1145/1416950.1416952.

Nguyen, D., Demeester, T., Trieschnigg, D., & Hiemstra, D. (2012). Federated search in the wild: The combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on information and knowledge management* (*CIKM 2012*), Maui, HI. doi:10.1145/2396761.2398535.

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2010*) (pp. 603–610), Geneva. doi:10.1145/1835449.1835550.

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, *43*(2), 531–548. doi:10.1016/j.ipm.2006.07.020.

Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., & Song, R. (2013). Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of the 10th NTCIR conference* (pp. 94–123), Tokyo.

Smucker, M. D., & Clarke, C. L. (2012). Modeling user variance in time-biased gain. In *Proceedings of the symposium on human–computer interaction and information retrieval* (*HCIR 2012*), Cambridge, CA. doi:10.1145/2391224.2391227.

Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q., & Orii, N. (2011). Overview of the NTCIR-9 INTENT task. In *Proceedings of the 9th NTCIR workshop meeting* (pp. 82–105), Tokyo.

Sormunen, E. (2002). Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of the 25th International ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2002*) (pp. 324–330), Tampere.

Turpin, A., Scholer, F., Jarvelin, K., Wu, M., & Culpepper, J. S. (2009). Including summaries in system evaluation. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2009*) (pp. 508–515), Boston, MA. doi:10.1145/1571941.1572029.

Vakkari, P., & Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. *Journal of the American Society for Information Science and Technology*, *55*(11), 963–969. doi:10.1002/asi.20046.

Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, *36*(5), 697–716. doi:10.1016/S0306-4573(00)00010-8.

Voorhees, E. M. (2001). Evaluation by highly relevant documents. In *Proceedings of the 24th international ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2001*) (pp. 74–82), New Orleans, LA. doi:10.1145/383952.383963.

Webber, W., Chandar, P., & Carterette, B. (2012). Alternative assessor disagreement and retrieval depth. In *Proceedings of the 21st ACM international conference on information and knowledge management* (*CIKM 2012*) (pp. 125–134), New York, NY. doi:10.1145/2396761.2396781.

Yilmaz, E., Shokouhi, M., Craswell, N., & Robertson, S. (2010). Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM international conference on information and knowledge management* (*CIKM 2010*) (pp. 1561–1564), Toronto, ON. doi:10.1145/1871437.1871672.

Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th International ACM SIGIR conference on research and development in information retrieval* (*SIGIR 2003*) (pp. 10–17), Toronto, ON. doi:10.1145/860435.860440

Zhou, K., Zha, H., Chang, Y., & Xue, G. R. (2014). Learning the gain values and discount factors of discounted cumulative gains. *IEEE Transactions on Knowledge and Data Engineering*, *26*(2), 391–404. doi:10.1109/TKDE.2012.252.