Exploiting User Disagreement for Web Search Evaluation: an Experimental Approach

Thomas Demeester[1], Robin Aly[2], Djoerd Hiemstra[2],
Dong Nguyen[2], Dolf Trieschnigg[2], Chris Develder[1]
[1] Ghent University - iMinds, Belgium
[2] University of Twente, The Netherlands
tdmeeste@intec.ugent.be, {r.aly, d.hiemstra}@utwente.nl
{d.nguyen, d.trieschnigg}@utwente.nl, cdvelder@intec.ugent.be

ABSTRACT

To express a more nuanced notion of relevance as compared to binary judgments, graded relevance levels can be used for the evaluation of search results. Especially in Web search, users strongly prefer top results over less relevant results, and yet they often disagree on which are the top results for a given information need. Whereas previous works have generally considered disagreement as a negative effect, this paper proposes a method to exploit this user disagreement by integrating it into the evaluation procedure.

First, we present experiments that investigate the user disagreement. We argue that, with a high disagreement, lower relevance levels might need to be promoted more than in the case where there is global consensus on the top results. This is formalized by introducing the User Disagreement Model, resulting in a weighting of the relevance levels with a probabilistic interpretation. A validity analysis is given, and we explain how to integrate the model with well-established evaluation metrics. Finally, we discuss a specific application of the model, in the estimation of suitable weights for the combined relevance of Web search snippets and pages.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Miscellaneous

Keywords

User disagreement, graded relevance, evaluation.

1. INTRODUCTION

Search engine evaluation uses relevance judgments of assessors to measure the effectiveness of search algorithms. Traditionally, a single assessor makes binary judgments about the relevance of documents based on a description of the information need at hand. Voorhees [22] showed that system comparisons are stable even for large disagreement among assessors. Because the research community agrees that relevance is in reality more complex than a binary label, modern evaluation measures consider graded relevance levels [18], groupings of documents into classes of the same type of content to evaluate the diversity of a ranking [25], and the ambiguity of search requests. Given the wide variety of users and internet content, we propose that in the Web context, it becomes increasingly questionable whether the judgments from one assessor lead to stable evaluation results. In this paper we investigate the disagreement of users on graded relevance levels, on two very different test collections: one with a large set of crowd-sourced Web search assessments, and one with less, but high-quality, judgments of both result snippets and result pages in a federated Web search setting.

Disagreement among users is an important aspect of search engine evaluation. Many works have investigated the influence of disagreement on search evaluation, (e.g., [4, 2, 22]). These works have been mainly based on binary relevance labels. The common approach to assess the disagreement is by measuring its amount, for example with the Jaccard coefficient or the kappa statistic, and to determine its influence by assessing the ranking of systems that the different judgment versions generate. Therefore, these works consider disagreement as a negative effect of including different judges, and test what the consequences of these unwanted effects are. In this paper we propose that disagreement should be seen as a reality that contains useful information, e.g., to estimate the relative importance of different levels of relevance. In Section 3 we will show that the overlap between judgments of top relevance for realistic Web data is rather low. Yet, Kekäläinen and Järvelin [15] propose that Web search evaluation should strongly favor documents of top relevance in the evaluation measure. This has been confirmed by Huang and Efthimiadis [12], as users tend to reformulate queries, rather than looking far down the result list. In that context, disagreement among users potentially has a strong effect on system evaluations. One of the key problems with integrating disagreement into evaluation measures is that there are no available models for disagreement. In this paper we propose one, referred to as the User Disagreement Model (UDM), for the disagreement in relevance judgments among users, and describe how the model's parameters can be accurately estimated with a limited number of queries judged by pairs of assessors.

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556268.

Current evaluation measures based on graded relevance either lack a motivation for the weights they are using, e.g., in the normalized Discounted Cumulative Gain (nDCG) measure [13], or, for probabilistically motivated models, the parameters are difficult to estimate, e.g., for the Graded Average Precision (GAP) measure [18]. We will use the UDM, based on the relevance opinion of multiple users, to make existing evaluation metrics for graded relevance more realistic.

In summary, this paper makes the following contributions to the field of search engine evaluation:

- We provide insight into user disagreement phenomena, with experimental results based on two test collections (Section 3).
- We propose a probabilistic model for user disagreement on relevance (Section 4).
- We show how our model can be used with existing graded relevance-base evaluation metrics to better reflect the real-world conditions of disagreeing users (Section 5).
- We indicate which parameters influence the robustness of our model (Section 6).
- We show how our model can be used to calculate suitable relevance weights for combined (snippet, page) relevance levels in Web search evaluation (Section 7).

2. RELATED WORK AND BACKGROUND

This section describes the relation of our disagreement model with other, well-established, aspects of search engine evaluation.

The evaluation of search engines usually involves modeling the following aspects: (i) the *label type* upon which the measure is based, (ii) a *user model* that describes how users process search results, (iii) *erroneous judgments* or *assessor bias*, and finally (iv) *evaluation metrics* that define the search effectiveness based on its input. We will now discuss these aspects in detail.

The *label types* for traditional evaluation metrics are limited to binary relevance, which is assumed to hold for all users having a well-defined information need. More recently, several works have improved upon this simplistic assumption, mainly into the following three directions: (i) capturing diversity by defining equivalence classes of documents [25, 7], (ii) capturing the ambiguity of queries by intent labels [1], and (iii) assuming that documents can have multiple grades of relevance [18, 13]. Compared to label types, the contribution in this paper is a different aspect of search engine evaluation. Instead of investigating how to assign different label types to documents that are valid over all users, our work acknowledges the fact that the labels naturally vary between users, and we define a model for user disagreement.

User models describe how users process ranked lists of search results. For example, Yilmaz et al. [23] proposed a probabilistic user model, where users read until reaching a random relevant document. The model by Moffat and Zobel [16] assumes that users stop reading a ranked list at a random rank. Such user models therefore also consider multiple users similar to our disagreement model. However, they focus on the reading behavior of users and assume that users agree on the same label for the same document. Our model is independent from the user's reading behavior and addresses the distribution of labels that different users might have. Another existing model that has a number of aspects in common with the UDM, is the user model underlying the GAP metric, by Robertson et al. [18], where one user is assumed to consider a result relevant, as soon as it has its relevance level at or above a specific cut-off level, with a specific distribution over the user population. However, estimating suitable collection-specific weights for these levels is not obvious, whereas it is automatically done by the UDM. Note that, where our model is simple and only based on relevance judgments, more complex user models have been proposed successfully, e.g., by Yilmaz et al. [24], based on the click data of real search sessions.

Investigating *judgment errors* and *assessor bias* is important, e.g., for assessors that are not trained or not motivated. In a recent study, Carterette and Soboroff [5] showed how different types of errors affect evaluation metrics and proposed strategies to compensate for errors, e.g., by selecting certain results for rejudging. These issues have in common with our model that they accept that the annotated labels are not necessarily 'true'. However, in contrast to previous research, we assume that the assessors are representative for the actual users and that an observed mismatch between annotation labels represents a form of user disagreement: we do not aim to "correct" it. For example, for highly erroneous assessments, the UDM-based relevance weights will be automatically adapted to large variations in relevance judgments for the same document. In an ideal scenario, however, judgment errors and bias would first be filtered from the assessments, upon which the UDM could provide suitable relevance weights for the remaining user disagreement.

The *evaluation metric* is a function defined on a label type and a user model and calculates the search effectiveness of a ranking to the current query. We see the main role of disagreement models in the adoption of strong existing evaluation measures, thereby providing relevance weights with a probabilistic interpretation, to incorporate the user disagreement in these measures.

In recent years, a lot of research on search evaluation has been done. Due to length constraints, we only mention a few more recent contributions that need to be explicitly mentioned in relation with the current paper. Some research on more advanced solutions than the traditional majority voting for aggregating multiple judgments has been done, e.g., by Hosseini et al. [11], who concurrently modeled the relevance of documents and the accuracy of assessors in a crowdsourcing setting. Our work instead models the disagreement among users, based on multiple judgments. Smucker and Clarke [20] already argued that it is essential for evaluation metrics to model variation between users, else the effect size of differences between retrieval systems would be overestimated. They proposed a suitable extension to the time-biased gain (TBG) metric. In relation to the combined snippet-based and page-based relevance weights given in Section 7, we acknowledge the work from Turpin et al. [21], on strategies to encorporate the snippet relevance into existing evaluation metrics.

3. EXPERIMENTAL INVESTIGATION OF USER DISAGREEMENT

In the following paragraphs, some properties of the relevance assessments for two different data collections will be investigated, focusing on issues related to the disagreement among assessors. The goal of this experimental analysis is to provide insights that are used to develop the user disagreement model (see Section 4).

3.1 Data

For the experiments presented in this paper, the relevance assessments for two publicly available datasets are used. The first is a crowd-sourced set of judgments for the TREC 2010 Relevance Feedback Track [3], and the second is a test collection composed in 2012 for research in federated Web search [17]. We will refer to these datasets as, respectively, the RelFeedback10 and the FedWeb12 data. Despite major differences, the relevance assessments for both datasets have three aspects in common, which are fundamental for the current work: (i) they both deal with Web data, (ii) use graded relevance levels, and (iii) contain independent relevance labels from different users. Some characteristics of the data, important for the remainder of this paper, are given below.

3.1.1 RelFeedback10 Data

The ground truth data for the RelFeedback10 data was created by means of mechanical Turk workers on English Web pages from the ClueWeb09 collection[1], for English search queries from the TREC 2009 Million Query Track [6]. Some of the documents also have prior 'gold' labels by NIST. The relevance levels are non-relevant (Non), relevant (Rel), and highly relevant (HRel). From the 20,232 judged results, we will focus on those 2,738 that contain 6 or more labels (making no distinction between the crowd and gold labels).

3.1.2 FedWeb12 Data

The FedWeb12 collection contains a large amount of sampled data from over a hundred diverse online search engines. It also contains relevance judgments by dedicated assessors, for both snippets and pages, of the first 10 results returned by each of these search engines. The test topics were taken from the TREC 2010 Web Track [8]. For the result snippets, the judgments for the perceived relevance (i.e., the estimate of page relevance, based on the snippet alone) are No, Unlikely, Maybe, and Sure, and for the pages, the levels are Non, relevant Rel, HRel, Key, and Nav. Only for some of the test topics (i.e., the navigational queries), the label Nav applied, and therefore in the current paper we merged it with the Key label. We will focus on the 15 test topics, that were entirely judged by two different assessors. More detailed information on the test topics, relevance judgments, and relevance distributions can be found in [9].

3.2 Tendencies in User Disagreement

The crowdsourced relevance judgments from the RelFeedback10 data are very noisy, such that there is a high disagreement among the assessors. For example, only 39% of the crowd assessments agree with the gold label. However, there is a trend toward higher labels for some documents, and lower labels for others, and this observation forms the starting point for the current paper. If we observe the label from one assessor for a particular result to be high, it can be expected that the other labels for that result tend to be high, too. We would like to quantify this effect, and therefore introduce $p_{\mathsf{HRel}|\bullet}$, the conditional probability of observing the highest label HRel, given the observation of a particular relevance label (denoted by $\bullet$) by another user. One of the goals of this section, is to show that an estimation of $p_{\mathsf{HRel}|\bullet}$ based on only 2 relevance labels per document, approximates a more precise estimate based on a higher number of labels per document.

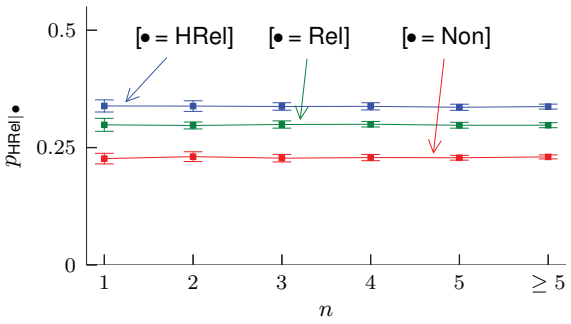

Figure 1: The probability $p_{\mathsf{HRel}|\bullet}$ of a HRel label, given an observed label $\bullet$ (for $\bullet$ = HRel, Rel, and Non), estimated from n judgments per document, showing the mean (± 1 std.) over all documents in the RelFeedback10 data with at least 6 labels.

In order to estimate, e.g., $p_{\mathsf{HRel}|\mathsf{Rel}}$, we conceived the following experiment. We went through all documents with at least 6 labels, randomly selected one of the labels, and if it was Rel, we calculated the fraction of HRel results among a random subset of n from the remaining labels. Averaging these values over the included documents, yielded our estimate of $p_{\mathsf{HRel}|\mathsf{Rel}}$. In order to verify the influence of which documents were included in the estimation (depending on the random choice of a single label), we repeated this procedure 50 times and calculated the mean and standard deviation of $p_{\mathsf{HRel}|\mathsf{Rel}}$. The results, with indication of the mean and one standard deviation above and below the mean, are shown in Fig. 1, with n varying from 1 to 5. The results are as intuitively expected: $p_{\mathsf{HRel}|\mathsf{HRel}} > p_{\mathsf{HRel}|\mathsf{Rel}} > p_{\mathsf{HRel}|\mathsf{Non}}$, and the estimates tend to be more precise (lower standard deviation) for increasing n. Yet, more importantly, the average is already fairly accurate even for $n = 1$: this is a support for our approach taken in Section 4, where we will only use double judgments to estimate similar parameters.

For IR evaluation purposes, finding a suitable choice for the weights of the different relevance levels is typically a problem. As mentioned before, Web search users are mostly interested in top results, i.e., the HRel results for the RelFeedback10 judgments. However, it seems that the weight for a result labeled Non should be almost as high as for a result labeled Rel, given that the average probability that a random other user would assign it the top label HRel, is not much lower. On the one hand, this demonstrates that care is required when relying on these noisy crowd judgments without further filtering. On the other hand, it leads to the following important insight: for a robust evaluation in the case of a large user disagreement, lower relevance levels might have to be promoted more than in the case where all assessors

[1] http://lemurproject.org/clueweb09/

would agree upon the top results. This will be formalized in Section 4, with the introduction of the user disagreement model.

3.3 User Disagreement vs. Intra-Judge Inconsistency

Consistency in terms of relevance assessments is often investigated with the purpose of asserting that a test collection is trustworthy for evaluation purposes. We actually want to use the extra information hidden in the user disagreement to our advantage, to obtain intuitive weights for the different relevance labels. In this section we analyze the FedWeb12 relevance judgments, focusing in detail on the consistency in terms of the highest relevance level, i.e., the Key results. We need both inter- and intra-assessor consistency to verify that the user disagreement is really disagreement among users, and distinguishable from inconsistency within a single user's assessments[2].

3.3.1 Overlap between users

Experiments done by Voorhees in [22] gave an average overlap in terms of relevant documents from two sets of binary relevance judgments ranging from 0.42 to 0.49, in the case of expert assessors judging topics on rather homogeneous datasets and a relevance cut-off level corresponding to our Rel level. However, at the top relevance level, different users tend to disagree more often. For the FedWeb12 judgments, the overlap for general Web search engines on Key results is only 0.36, and even lower for the other verticals. In fact, the level up to which the assessors agree can be used to estimate the practical upper limit of precision and recall on the performance of retrieval systems that are evaluated with the assessed test topics. For the TREC-4 data [22], there appeared to be a limit of 65%, whereas it is below 50% for Key relevance on the FedWeb12 data. For a less strict relevance level like Rel (see [9]), the results on the FedWeb12 data appeared comparable to those reported by Voorhees, but then again, in the Web search context, users strongly prefer top relevance.

3.3.2 Self-consistency vs. cross-user consistency

In the following paragraphs, we will demonstrate for the (high-quality) FedWeb12 relevance assessments, that variations in opinion between users dominate the inconsistencies in the judgments. Actually, for this type of relevance judgments we can identify at least three important sources of inconsistencies: (i) judgment errors, (ii) uncertainty for each judge in assigning relevance (also called intra-judge reliability, see [19]), and (iii) inter-user disagreement (elsewhere in this paper shortly called user disagreement), reflected by inter-judge differences. The least important type are plain mistakes. These happen at random with all assessors, and can be due to a sudden drop in concentration, hitting the wrong key when assigning a specific relevance level, etc. However, we observe inconsistencies between judgments from the same assessors, with variations that are too large to stem solely from such random mistakes. Users especially seem to inconsistenly assign different adjacent relevance levels, in the case of graded relevance assessments. In the following paragraphs, we will answer two research questions. How important is the intra-judge inconsistency with respect to inter-user disagreement, and can we make a distinction in user disagreement between different relevance levels despite the intra-judge inconsistencies?

[2]However, the *origin* of that user disagreement (e.g., different background knowledge, or even a completely different intent), does not affect the validity of the user disagreement model, introduced in Section 4.

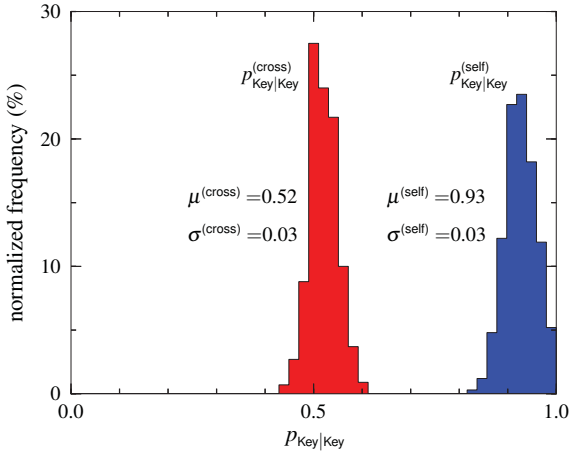

Figure 2: Normalized histograms of the probability for results labeled Key in one judgment to receive label Key in another, by a different user ('cross'), respectively, the same user ('self').

The first question we address is to establish self- vs. cross-user inconsistencies. We devised an experiment to obtain a better view on the influence of the self-inconsistency on the apparent difference in opinions between assessors, based on the FedWeb12 data. The resources used for FedWeb12 include several large Web search engines, implying an important overlap in their results for the test topics. From those test topics entirely judged by two assessors, we selected the pages corresponding to those URLs that appeared at least two times per topic, and whose multiple occurences were judged independently by the same judge. We found on average around 3 independent judgments by each of both users for the 387 selected URLs. We used a bootstrap sampling setup to visualize the variance in labels over multiple judgments by the same user for the same result. Assuming that each of the judgments effectively given by one user to a particular result is equally likely, we sampled the sets of judgments to generate single judgment lists for both of the users, which we then used to estimate the parameters $p_{Key|Key}$. A first result is shown in Fig. 2. It shows normalized histograms over 1000 such samples for the probability $p_{Key|Key}$ that Key relevance is assigned in one assessment, given Key relevance from another. The case where both assessments come from the same user, has an average probability $\mu^{(self)}$ = 0.92, whereas for different users we get $\mu^{(cross)}$ = 0.52. The standard deviation of the histograms reflects the self-inconsistency of one assessor (including possible random errors), and the separation between both histograms corresponds to the difference in opinion between different assessors. Clearly this difference in opinion between users overshadows the uncertainty a single user has.

The second issue to be addressed is whether the assessors' self-inconsistency does not prevent us from making a clear distinction between parameters $p_{Key|\bullet}$ for different relevance levels $\bullet$. Figure 3 shows histograms for $p_{Key|Non}$, $p_{Key|Rel}$, and

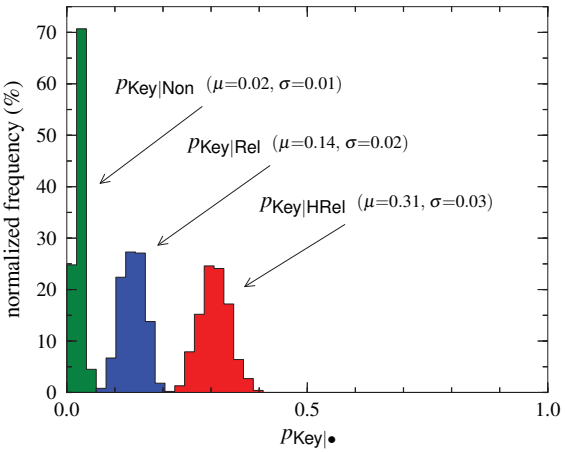

Figure 3: Influence of self-inconsistency on $p_{\text{Key}|\bullet}$ (for $\bullet$ =HRel, Rel, and Non), the probability that one assessor judges a result Key, given that another assessor judged it with another relevance level.

$p_{\text{Key}|\text{HRel}}$, based on the same set of samples as used for Fig. 2, indicating μ, the mean value of $p_{\text{Key}|\bullet}$, and its standard deviation σ. These μ and σ lead us to observe that, despite the assesors' self-inconsistencies, we can distinguish the various $p_{\text{Key}|\bullet}$ probabilities very well from each other.

Note that the $p_{\text{Key}|\text{Non}}$ estimates are very low for the FedWeb12 data. Hence, we propose to neglect them, when we use the parameters $p_{\text{Key}|\bullet}$ in the context of the UDM (see the next section). The validity of this approximation is further motivated by the fact that the estimated $p_{\text{Key}|\text{Non}}$ from our data is an overestimation (related to the fact that only a particular subset of the non-relevant pages were judged, i.e., those with snippet label above No, see [9] for details). Most likely some of the few cases where one user assigns Key relevance to a result that another finds completely non-relevant, are due to the aforementioned random mistakes. Clearly, we do not want that random error to impact global relevance metrics, where large amounts of non-relevant results would erroneously contribute a small but non-neglible amount of relevance (proportional to the non-zero $p_{\text{Key}|\text{Non}}$).

4. USER DISAGREEMENT MODEL

This section describes the proposed User Disagreement Model (UDM). Based on the experiments from Section 3, we first outline the context and goals for the model. Then we provide a somewhat naive illustration to explain the intuition behind UDM. This is followed by the description of the model itself, together with a verification using the RelFeedback10 data. Finally, we discuss certain subtleties of the model, specific for the FedWeb12 data.

4.1 Context and Goal

The UDM is proposed in a Web search context, where we adopt the notion that many users strongly prefer results of the highest relevance level, and are no longer satisfied with partially relevant results, from Kekäläinen and Järvelin [15] and Huang and Efthimiadis [12]. However, there is a significant disagreement on top relevance among assessors, as illustrated by the RelFeedback10 data and the FedWeb12 data in Sections 3.2, respectively, 3.3. It was also shown for the FedWeb12 data that this disagreement among assessors is more important than the uncertainty from a single assessor. We hence propose that the inconsistency among assessors is a *fundamental* difference, also present among actual users (assuming the assessors are representative for these users). The reason for the inter-assessor disagreement can be found partly in the insufficient description of the information need. However, even with a very detailed and unambiguous description, there would always be a difference in opinion, due to a different background knowledge, education, culture, personal taste, etc. Note that, for assessments with a very large variation in relevance (like the RelFeedback10 data), the observed disagreement might be mainly determined by the noisy judgments. However, the validity of the UDM model does not depend on the actual source of disagreement.

The main goals of the UDM are the following. We would like to have an evaluation system that is robust with respect to variations in assessments because of the aforementioned user disagreement. Indeed, search engines would prefer to predict the top-relevant documents for the complete user base, not just a single user (the assessor). Thus, we want to infer what documents would be considered a top result by, e.g., at least 1 out of 4 users, even though we only have, e.g., a single assessor's judgments available.

In the context of evaluation, we designed the UDM as a means to match the weights from graded relevance metrics to the collection under evaluation and the intended users. For example, if relevance assessments obtained by crowdsourcing are highly variable, the relevance levels below top might need a higher weight, in order to fairly take into account those results that would be judged top by many other users, but not the considered assessor. Alternatively, for relevance assessments based on highly precise information needs, the disagreement would be lower, and the relevance levels below top should be given a lower impact, because almost no users with such needs would consider them top.

4.2 Intuition

Consider the simple (artifical) configuration sketched in Fig. 4, where a collection of 8 documents is shown with the relevance opinions by 4 users over 4 relevance levels (indicated with different colors) for a particular query. We assume that each user is only interested in results with the top relevance level (indicated with a thick box), in this case corresponding to the relevance label Key. The figure shows a strong variation in assigned relevance grades among the users, which we refer to as the user disagreement. Now, assume that we only have one assessor (indicated with an asterisk), user 3, from whom we can observe the relevance labels. However, if we only return document $d2$ (considered top by the assessor), the other users are not satisfied. Yet, if we are aware of existing user disagreement in general (i.e., not for the specific given query), we could try to please them all, e.g., by returning documents at least one of them considers top. For each of the judgments from the assessor, we want to estimate the probability that one or more users would find the corresponding document a top result. We observe that for results labeled HRel, that probability can be estimated as 1/2 (because in half of the cases where the assessor indicated HRel, at least one of the others labeled it with Key), and for those labeled Rel (slightly relevant), it is only 1/3. As seen from Fig. 4, no one assigns top relevance to

Figure 4: Illustration of user disagreement for a specific query, with the different colors indicating the graded relevance levels, and the thick black boxes the top results per user.

a result if another would label it Non, and the corresponding probability is 0. For results labeled Key by the assessor, we already know at least one of the users consider it top, hence the corresponding probability is 1. We can use these probabilities as weights for the respective relevance levels. Going over all the labels from the assessor, we can add these up to count the *effective* number of top results, which turns out to be 4, the number of documents considered top by at least one judge. In order to estimate the required probabilities, we used the relevance opinion of all other users, which in reality are not known. It will turn out that, if we assume all users to have a similar disagreement amongst them, we only need to know which documents one other user considers top.

Now that we have outlined the basic rationale of the UDM, we will more formally address the model details.

4.3 UDM Description

Consider a group of users that each want to retrieve top results, but the individual opinions on what results are top, vary amongst them. We define the total set of results considered relevant according to the UDM, as those that would receive the top relevance label T by at least M out of N users (for example, at least one out of three). The parameter that defines the UDM is the probability that a document with one observed label i ($i = 1, \ldots, T$) is relevant in that sense, where the assessor belongs to the N considered users. More specifically, this parameter is written $P_{T|i}^{(M/N)}$, and denotes the probability that at least M out of N randomly selected users consider it a top result, given the relevance label i from one of these N users. We call $P_{T|i}^{(M/N)}$ the UDM-based *relevance weight* for level i.

We consider M and N to be design parameters, depending on the purpose of the evaluation. For example, for small M and large N, the effective number of top results (i.e., according to the UDM) would be higher than the number of top results indicated by a single user, smoothing out differences in opinion among users and effectively leading to a more robust and less strict relevance criterion. The relevance weights can for instance be used to calculate the expected number of relevant results in a result list (with the UDM notion of relevance introduced above), or as the probability of relevance for each result (again, following the UDM), which allows creating an ideally ranked result list as a reference for evaluation.

4.4 Estimation of UDM Relevance Weights

Consider the basic scenario with two users, A (the assessor) and R (a random user), and the case $M = 1$ and $N = 2$. For each result, we want to estimate the probability that at least one of these two users labels it top, given the assessor label i_A,

$$P_{T|i}^{(1/2)} = Pr\left[i_A = T \vee i_R = T | i_A = i\right]. \quad (1)$$

It turns out that

$$\begin{aligned} P_{T|i}^{(1/2)} &= 1, & i = T, \\ &= p_{T|i}, & i < T. \end{aligned} \quad (2)$$

where i_R is the relevance label assigned by random user, and $i < T$ denotes that level i is below the top level. The parameter $p_{T|i}$ is the probability that we will observe the label T from one user, if we observed the label i from another user, as introduced and investigated in Section 3.2. As shown in Fig. 1, this parameter can be estimated accurately using judgments from only two assessors. Other dependencies, e.g., on the test topics, will be discussed in Section 6.2.

The result described above is trivial, but it can be used to describe more general cases. First, consider the case of $M = 1$, and $N \geq 2$, where we want to estimate the probability that at least one among N users judges the considered document top. We find

$$\begin{aligned} P_{T|i}^{(1/N)} &= 1, & i = T, \\ &= 1 - (1 - p_{T|i})^{N-1}, & i < T, \end{aligned} \quad (3)$$

For the case $i < T$, we assumed the same disagreement behavior and independence between the users. The probability that at least one out of N users considers a result top, given that one of them already assigned level i below top, is the complement of the probability that the $N-1$ remaining users also consider it below top, and directly leads to (3). Before turning to the more general case, (3) allows making the following observations on the behavior of the relevance model. The probability that at least one out of N users assigns top relevance to a result judged lower by the assessor, goes to one if N is large enough and $p_{T|i} > 0$ (which means there is disagreement among the users, and the considered relevance level is above total non-relevance). If the different judges would always agree on the top relevance level, $P_{T|i}^{(1/N)}$, ($i < T$) would go to zero, and the relevance model would become binary relevance on the top level. These observations agree with our intuition and the prescribed idea of the UDM.

In the most general case, with $0 \leq M \leq N$, we find

$$P_{T|T}^{(M/N)} = \sum_{m=M-1}^{N-1} \binom{N-1}{m} (p_{T|T})^m (1 - p_{T|T})^{N-1-m}, \quad (4)$$

$$P_{T|i}^{(M/N)} = \sum_{m=M}^{N-1} \binom{N-1}{m} (p_{T|i})^m (1 - p_{T|i})^{N-1-m}, \quad i < T \quad (5)$$

These formulas can be derived by summing the probabilities for each allowed configuration (with M or more top results, taking into account the observed label), over all possible combinations of the labels. Alternatively, they can be written down directly from the binomimal cumulative distribution, with the Bernoulli parameter $p_{T|i}$. Using the binomial theorem, it can be easily verified that for the case $M = 1$, expression (5) leads to (3), and (4) becomes 1.

One property of $P_{T|T}^{(M/N)}$ should be mentioned here. If $M > 1$ and there is no perfect agreement on top relevance, $P_{T|T}^{(M/N)} < 1$. This means, for instance, that a result assigned top relevance by the assessor would contribute less than 1

Table 1: Comparison between experimental values for $P^{(M/N)}_{\mathsf{HRel|Rel}}$, and those predicted with eq. (5), using the RelFeedback10 data.

	$P^{(M/N)}_{\mathsf{HRel|Rel}}$	
M / N	experimental	prediction
1 / 3	0.49	0.51
2 / 3	0.10	0.09
2 / 4	0.23	0.21
2 / 5	0.35	0.35

to the expected number of relevant results, according to the UDM relevance model with this particular choice of parameters M and N. This leads to an effectively even stricter relevance cut-off than the highest relevance level T. This may be undesirable in some cases. For example, the expected precision for a set of results all assessed to be top relevant would be below 1, in which case the GAP metric (see Section 5) could not be used for evaluation. By conditioning on more than a single assessment, this problem could be avoided, but in that case, the model would require more than 2 judgment sets for the training topics. Note that the choice of using a single observed relevance judgment in the definition of the UDM is no inherent limitation, although it keeps the model simple. However, we assume that only for a few test topics, we can afford to gather double judgments[3]. For a more complicated model, e.g., with 2 observed labels, the number of parameters to be estimated becomes the square of the number of relevance levels, and too many three-fold judgments would be required to obtain good estimates for these. We therefore focus on the case $\alpha = 1/N$, $N \geq 2$.

4.5 Verification

Since the RelFeedback10 relevance assessments contain multiple labels per document, we can easily verify the validity of the prediction formulas for the relevance weights. During the experiment on the RelFeedback10 data described in Section 3.2, we recorded the fraction of cases in which at least M out of the N considered labels were HRel. The resulting estimates for $P^{1/N}_{\mathsf{HRel|Rel}}$ and $P^{1/N}_{\mathsf{HRel|Non}}$ are shown as rectangular markers in Fig. 5, with indication of the error (plus or minus one standard deviation) depending on the chosen sample of documents. The full lines and dotted lines respectively show the values predicted by (3), and the potential error due to inaccuracies in the estimation of $p_{\mathsf{HRel|Rel}}$, respectively, $p_{\mathsf{HRel|Non}}$. Note that these predictions are only valid for integer values of N, but were drawn as continuous lines for clarity. There is a clear correspondence, although the experimental values for smaller fractions $1/N$ are slightly lower than the predicted behavior. As a verification of the general formula 5, Table 1 provides a comparison between the experimental and the predicted values for different M and N, with again a good correspondence.

[3]Alternative strategies are possible, e.g., by using crowdsourcing techniques to harvest large amounts of cheap but noisy judgments, and filtering out the least trustworthy ones.

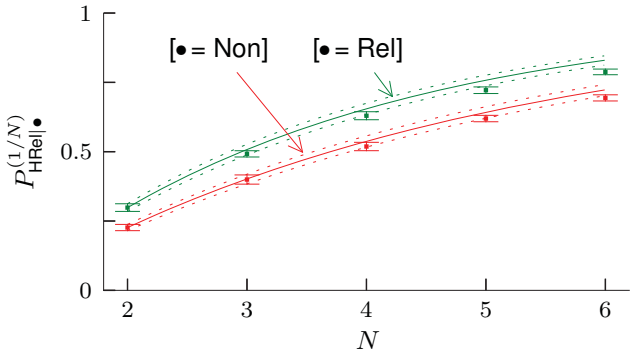

Figure 5: Comparison between experimental values for $P^{(1/N)}_{\mathsf{HRel|Rel}}$ and $P^{(1/N)}_{\mathsf{HRel|Non}}$ (as the mean ± 1 std., with markers), and those predicted with eq. (3) (full lines), using the RelFeedback10 data.

5. GRADED RELEVANCE METRICS AND THE UDM

This section describes possible ways to incorporate the UDM into the existing graded relevance evaluation metrics nDCG and GAP.

nDCG is a flexible measure, in the sense that it allows the user to define a suitable gain function for weighting the relevance level i_k at position k, and a discount function that reflects how fast the contribution of relevant items ranked lower in the result list should decrease. In general, it can be written as follows

$$\mathrm{nDCG}^{(\mathrm{gain})}_{\mathrm{discount}} = \frac{1}{Z} \sum_k \mathrm{gain}(i_k)\,\mathrm{discount}(k) \qquad (6)$$

where Z is a normalization constant that ensures a value of 1 for the ideal ranking. A popular gain function is the exponential $(2^{i_k} - 1)$, which promotes the highest relevance levels, and a typical discount function is the the logarithmic discount $1/\log_2(k+1)$ or the zipfian discount $1/k$. Further information can be found in [14]. These functions are well-chosen heuristics, and we propose to use the relevance weight $P^{(1/N)}_{i_T|i_k}$ of the result at rank k as $\mathrm{gain}(i_k)$. Note that the measure still remains inherently heuristic.

GAP is a highly informative and discriminative measure, which has a theoretical foundation and a nice probabilistic interpretation. However, because of these properties we must be careful when incorporating the UDM. The user model on which GAP is based, assumes that each user has a fixed cut-off relevance level, and considers results from any level equal to or higher than this cut-off level as relevant. The relevant parameters for the model are q_i, the probability that a user finds a result with relevance level i or higher relevant[4], whereby $q_T = 1$ (for the top relevance level T).

[4]The original formulation by Robertson et al. [18] introduces the fundamental parameters g_i, the probability that a user finds *only* results with relevance level i or higher relevant, which form a probability distribution over the space of users (with $\sum_{i=1}^{T} g_i = 1$). The parameters q_i introduced above form the corresponding cumulative distribution function, $q_i = \sum_{j=1}^{i} g_j$, and are more convenient to make the link with the UDM.

For a ranked list of K results, GAP is now defined as

$$\text{GAP} = \frac{\sum_{k=1}^{K} \frac{1}{k} \sum_{\kappa=1}^{k} q_{\min(i_\kappa, i_k)}}{\sum_{i=1}^{T} R_i \, q_i} \tag{7}$$

in which R_i is the number of results with relevance level i in the result list. It can be shown that the denominator of this expression represents the expected number of relevant results in the list, based on the probabilistic user model. In the special case where all users would have their relevance cut-off at the highest level ($q_i = 0, \ \forall i < T$), the denominator becomes the number of top results, and GAP becomes the average precision measure based on binary relevance at highest level T.

To our knowledge, the UDM cannot be directly translated into the user model underlying GAP, because the former is defined over different relevance opinions for a *result*, whereas the latter is based on varying user opinions for each *relevance level*. Nevertheless, all required parallels are there to apply the UDM to GAP. To do so, we only need to replace q_i in (7) by the UDM relevance weights $P_{T|i}^{(1/N)}$. The denominator then denotes the expected number of top results according to the probabilistic UDM, and if all users had the same opinion, the GAP would also become the average precision at the top level. A similar derivation of GAP as in [18] could be done based on the UDM (whereby the cumulative user disagreement probability distribution to start from would be the ordered series of $P_{T|i}^{(1/N)}$ parameters), but would be somewhat artificial and is left out due to length restrictions. We however propose that the interesting fundamental properties of GAP remain valid even when based on the UDM.

6. ROBUSTNESS OF THE UDM

6.1 UDM-based Evaluation

We propose that UDM-based evaluation is more robust than when based on a single set of relevance judgments, and motivate it with following experiment. We again consider the double-judged test topics from the FedWeb12 dataset. The evaluation is considered robust, if two different relevance judgment sets for a particular test topic yield similar results. We use a leave-one-out cross-validation setting, whereby we present the average and standard deviation of a number of metrics over each of the test topics, based on the parameters estimated from the others. For the considered test topic, we used one judgment set as the reference, and the other as the 'retrieved' set (creating a ranked set by ranking according to descending relevance levels). The first metric considered is average precision (AP), based on binary Key relevance. We also consider $\text{GAP}^{(1/N)}$ for N values of 2, 3, and 4. Finally, we report $\text{nDCG}_d^{(g)}$ for different gains g and discount functions d. For the gain, (exp) denotes the traditional exponential gain, whereas $(1/N)$ means the coefficients $P_{\text{Key}|\bullet}^{(1/N)}$ are used as gain function. As discount functions, we report results for the logarithmic $(1/\log_2 n)$ and zipfian $(1/n)$ case. The results are shown in Table 2. The listed metrics would all be one if both judges had agreed upon Key relevance. A lower value indicates a larger difference between both baselines for the considered metric, and will lead to a larger dependency on the particular baseline when the metric is applied to a candidate ranking. The AP metric appears to be the least robust. The reported AP value of 0.48 means that an AP of 0.48 is the maximum value that can be used to compare retrieval algorithms (if one wants to judge on the strict notion of Key relevance), because the judges only agree up to this level. Whereas the AP corresponds to taking a single user's opinion into account, the GAP results (which consider those results relevant that would be judged Key by at least one out of N users) increase for higher N, leading to a more robust evaluation. This effectively means that the below-top relevance levels get a higher weight. A similar behavior is observed for the nDCG measures $\text{nDCG}_{\log}^{(1/N)}$, based on the same weights. The fact that this metric appears to be much more robust than GAP, is due to the logarithmic discount function, where exchanging a Key result in the top with a lower result on a lower rank is penalized less heavily than for nDCG with the zipfian discount function, or GAP.

Table 2: Mutual evaluation of paired judgment sets with different evaluation metrics (mean ± 1 std.).

AP	0.48 ± 0.28
$\text{GAP}^{(1/2)}$	0.65 ± 0.18
$\text{GAP}^{(1/3)}$	0.69 ± 0.16
$\text{GAP}^{(1/4)}$	0.71 ± 0.16
$\text{nDCG}_{\text{zipf}}^{(\text{exp})}$	0.70 ± 0.23
$\text{nDCG}_{\log}^{(\text{exp})}$	0.86 ± 0.09
$\text{nDCG}_{\log}^{(1/2)}$	0.84 ± 0.11
$\text{nDCG}_{\log}^{(1/3)}$	0.87 ± 0.09
$\text{nDCG}_{\log}^{(1/4)}$	0.89 ± 0.07

6.2 Parameter Dependencies

In previous sections we assumed that the parameters $p_{T|i}$ are constant over the considered collection. For the RelFeedback10 experiments from Section 3.2, they were calculated by sampling over all documents. Sometimes, however, better choices should be made due to the inhomegeneous character of the test collection, as is the case for the FedWeb12 data. Two sources of inhomogeneity will be shortly discussed in the following paragraphs: the type of test topics, and the origin of the results.

For test topics with an information need description in the sense of '*The user is looking for information about ...*' (as is mostly the case for the topics used in the FedWeb12 judgments), the $p_{i_T|i_t}$ will probably be higher than for cases where the information need is described in finer detail. The test topics were originally designed for the TREC 2010 Web Track that focused on result diversity, and can be divided into ambiguous topics (with multiple distinct interpretations), or faceted (with different possible aspects of the same interpretation), see [8]. Among our twice judged topics, 7 are ambiguous and 8 are faceted. The comparison of the parameters $p_{\text{Key}|\bullet}$ estimated separately for the different types of topics, see Table 3 shows that for the ambiguous queries there is a higher chance that assessors strongly disagree ($p_{\text{Key}|\text{Rel}}$ is larger), whereas for the faceted topics the distinction between both highest relevance levels is less clear ($p_{\text{Key}|\text{HRel}}$ is higher). In this paper, we do not make any further distinction between both types of topics and use the parameter estimates for all test topics together. However, different parameter sets could be used for evalution scenar-

Table 3: Dependency of the parameters $p_{T|i}$ on the test topics (ambiguous vs. faceted), result list quality, and result ranks (top 5 vs. top 10).

# results	top 5		top 10		top 10	
test topics	all		all		ambig.	facet.
high-quality	no	yes	no	yes	yes	
$p_{\mathsf{Key}\|\mathsf{Rel}}$	0.03	0.16	0.02	**0.15**	0.18	0.13
$p_{\mathsf{Key}\|\mathsf{HRel}}$	0.09	0.25	0.06	**0.23**	0.22	0.26

ios where a distinction is needed. In future work, we will study the influence of the test topics with the much larger FedWeb13 test collection, for which the test topics are less homogeneous (see Demeester et al. [10]).

The judged results can be grouped as lists of 10 results per query and per resource (i.e., the top 10 results for that query from each of a wide variety of search engines). A large fraction of the search engines produced only non-relevant results for most queries. However, some search engines (e.g., general Web search engines) provided high-quality result lists. We can, for example, expect the following: when we only consider the low-quality resources, the probability that one user would give a Key label for a result labeled Rel by another, will probably be much lower than when we only look at result lists from strong search engines. Using an overall average for $p_{T|i}$ would hence lead to a value effectively too low for the test results of interest (i.e., those that originate from potentially interesting resources). In our experimental setting, we made a distinction between low- and high-quality result lists. Those with at least one Key result are considered high-quality lists, and the others low-quality. Table 3 displays $p_{\mathsf{Key}|\mathsf{Rel}}$ and $p_{T|i}$, calculated separately for both groups of result lists (indicated with 'high-quality' in the table). The difference is very large, as expected. For the current paper, we decided to only use the high-quality result lists to estimate these parameters (indicated in bold in Table 3). This implicates that the evaluation of the higher relevance grades is more accurate, whereas the over-estimated influence of the lower relevance levels is still small. For example, with GAP, the influence of lower ranked results is discounted rapidly, and therefore the potentially long tail of results with a low relevance hardly contributes to the final score, despite an over-estimated weight of $p_{\mathsf{Key}|\mathsf{low}}$.

For other search scenarios, other choices are possible. For example, for a single ranked result list in general (i.e., not in the federated search context), one might consider only the top-k results for all doubly judged test queries, especially in combination with a metric that is calculated up to a rank k. As a demonstration that the considered position in the ranked list does influence our parameters, we estimated them from the top-5, respectively from the top-10, of each result list (see Table 3). The estimated $p_{i_T|i_t}$ values are consistently but only moderately higher for the lists with the higher average quality, i.e., the top-5 lists.

The above analysis has brought up some clear difficulties in estimating the parameters in a consistent way. It should be noted that most of these issues could be translated into similar issues within the probabilistic user model that Robertson et al. [18] proposed, where the required parameters are assumed constant as well.

7. APPLICATION

In Demeester et al. [9] it is demonstrated that the perceived relevance based on the snippet is often not in line with the actual page relevance, and that ideally a Web IR system should be evaluated based on the relevance experienced by the user, when the snippet and page relevance are combined. This means, e.g., that top pages behind bad snippets should be penalized when ranked at the top. One straightforward way to do that, is by creating combined relevance levels consisting of all combinations between relevance levels for pages and snippets. Examples are (Maybe, HRel), or (Sure, Rel), where we first mention the snippet level, followed by the page level. The experienced top relevance level in this context is defined as the combination of the top levels for snippets and pages separately, i.e., (Sure, Key).

A direct application of metrics like nDCG or GAP is difficult, because these would require an ordering according to relevance of these combined relevance levels. Yet, from the viewpoint of the user, should (Maybe, HRel) be the higher label, or rather (Sure, Rel)? We apply the UDM to find relevance weights with a probabilistic interpretation. The results are shown in Table 4, for $M/N = 1/3$. Each row presents a possible combined relevance level, and the rows are ordered for decreasing values of $P^{(1/3)}_{(\mathsf{Sure},\mathsf{Key}|\mathsf{L}_{\mathsf{snippet}},\mathsf{L}_{\mathsf{page}})}$, where $\mathsf{L}_{\mathsf{snippet}}$ and $\mathsf{L}_{\mathsf{page}}$ denote the different snippet, respectively, page relevance levels. For example, $P^{(1/3)}_{(\mathsf{Sure},\mathsf{Key}|\mathsf{Maybe},\mathsf{HRel})} = 0.31$ means that there is a 31% chance that at least one out of three users would assign the top relevance level (Sure, Key) to a result, given that we observed one of them assigning the snippet label Maybe and the page label HRel. For comparison, the relevance weights based on page relevance alone, are given as well. It appears that a snippet judgment below Sure leads to a degradation of the combined relevance, with respect to the corresponding relevance weight for the page relevance alone.

There are many combined relevance levels, and given a limited number of twice judged topics, we can therefore doubt the accuracy of all calculated parameters. For example, we only found 36 cases where one assessor judged a result as (No, HRel), and only in five of those cases the other assessor assigned top relevance to the same result. The result of 0.26 for $\alpha = 1/3$ is therefore most likely a poor estimate of the actual probability. Those relevance levels for which less than 50 cases were present in the data (an arbitrarily chosen boundary), are indicated between brackets, as being not accurate. However, the less accurate such an estimate becomes, the less often these cases appear in the collection, and the less influence possible variations of the estimates have on global evaluation scores. For the cases that often occur, the estimate is more accurate, and these most strongly influence the evaluation results. For example, there were 372 occurences of (Maybe, HRel) in our twice judged topics, with 87 matching top judgments by the other judge.

8. CONCLUSIONS

We introduced the probabilistic User Disagreement Model, which models fundamental differences in opinion on relevance among users. Empirical results were given based on two recent test collections (the TREC'10 Relevance Feedback judgments, and the FedWeb12 collection), and we provided insights into issues related to user disagreement, such

Table 4: Relevance weights for combined (snippet, page) relevance levels, and for page levels alone (with brackets indicating where less than 50 judgments were found).

$L_{snippet}$	L_{page}	$P^{(1/3)}_{(Sure,Key\|L_{snippet},L_{page})}$	$P^{(1/3)}_{Key\|L_{page}}$
Sure	Key	1	1
Maybe	Key	0.42	
Sure	HRel	0.41	0.41
Unlikely	Key	(0.38)	
No	Key	(0.34)	
Maybe	HRel	0.31	
Sure	Rel	0.29	0.28
No	HRel	(0.26)	
Maybe	Rel	0.20	
Unlikely	HRel	(0.13)	
Unlikely	Rel	0.13	
No	Rel	0.12	
(any)	Non	0	0

as the influence of intra-assessor inconsistencies, and the robustness of the UDM-based probability estimations. It was also shown how the model could be incorporated into the graded relevance metrics nDCG and GAP, as such allowing these measures to take into account user disagreement. As an illustration, the model was used to estimate the required relevance weights for the combined evaluation of snippet and page relevance in a federated Web search context.

This paper's main contributions are theoretical, and in future contributions we will reap the benefits in more tangible results, by evaluating actual retrieval results.

9. ACKNOWLEDGMENTS

This work was funded by Ghent University - iMinds in Belgium, by the EU Project AXES (FP7-269980), the Netherlands Organization for Scientific Research (NWO), grants 639.022.809 and 640.005.002 (FACT), and the Dutch national project COMMIT.

10. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*, pages 5–14. ACM, 2009.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR '08*, pages 667–674. ACM, 2008.

[3] C. Buckley, M. Lease, and M. D. Smucker. Overview of the TREC 2010 Relevance Feedback Track. In *TREC*, 2010.

[4] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5):619 – 627, 1992.

[5] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *SIGIR'10*, pages 539–546. ACM, 2010.

[6] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million Query Track 2009 Overview. In *TREC*, 2009.

[7] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *ICTIR'09*, volume 5766, pages 188–199. Springer, 2009.

[8] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. *TREC*, pages 1–9, 2010.

[9] T. Demeester, D. Nguyen, D. Trieschnigg, C. Develder, and D. Hiemstra. What Snippets Say about Pages in Federated Web Search. In *AIRS'12*, pages 250–261, 2012.

[10] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the TREC 2013 Federated Web Search Track. *TREC*, 2013.

[11] M. Hosseini, I. J. Cox, N. Milić-frayling, G. Kazai, and V. Vinay. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. *ECIR'12*, pages 182–194, 2012.

[12] J. Huang and E. N. Efthimiadis. Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. *CIKM'09*, pages 77–86. ACM, 2009.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20 (4):422–446, 2002.

[14] E. Kanoulas and J. A. Aslam. Empirical Justification of the Gain and Discount Function for nDCG. In *CIKM'09*, pages 611–620, 2009.

[15] J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[16] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, 2008.

[17] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated Search in the Wild: the Combined Power of over a Hundred Search Engines. In *CIKM'12*, pages 1874–1878. ACM, 2012.

[18] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending Average Precision to Graded Relevance Judgments. In *SIGIR'10*, pages 603–610, 2010.

[19] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *SIGIR'11*, pages 1063–1072. ACM, 2011.

[20] M. D. Smucker and C. L. A. Clarke. Modeling User Variance in Time-Biased Gain. *HCIR'12*, 2012.

[21] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including Summaries in System Evaluation. *SIGIR'09*, pages 508-515. ACM, 2009.

[22] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[23] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR'08*, pages 587–594. ACM, 2008.

[24] E. Yilmaz, M. Shokouhi, N. Craswell, and S. E. Robertson. Expected browsing utility for web search evaluation. *CIKM'10*, pages 1561-1564. ACM, 2010.

[25] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR'03*, pages 10–17. ACM, 2003.