Modelling Randomness in Relevance Judgments and Evaluation Measures

Marco Ferrante[1], Nicola Ferro[2], and Silvia Pontarollo[1]

[1] Department of Mathematics, University of Padua, Italy
{ferrante,spontaro}@math.unipd.it
[2] Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it

Abstract. We propose a general stochastic approach which defines relevance as a set of binomial random variables where the expectation p of each variable indicates the quantity of relevance for each relevance grade. This represents the first step in the direction of modelling evaluation measures as a transformation of random variables, turning them into random evaluation measures. We show that a consequence of this new approach is to remove the distinction between binary and multi-graded measures and, at the same time, to deal with incomplete information, providing a single unified framework for all these different aspects. We experiment on TREC collections to show how these new random measures correlate to existing ones and which desirable properties, such as robustness to pool downsampling and discriminative power, they have.

1 Introduction

Relevance judgements are at the core of *Information Retrieval (IR)* evaluation since they determine and inform all the subsequent scoring and comparison of IR systems. For this reason, over the years, a lot of effort has been put in their creation and in ensuring their quality, see e.g. [7, 20], also in a crowd-sourcing context [1].

We know that relevance assessment is a not deterministic process, as witnessed by different studies on inter-assessor agreement [16, 17] and as exploited by algorithms to merge relevance labels in a crowd-sourcing context [5]. However, once relevance judgments have been created – either by traditional assessors or with sophisticated algorithms merging labels from crowd-assessors – we seem to forget their intrinsic randomness and we consider them as if they were deterministic: for example, evaluation measures just handle the relevance judgment associated with a document as exact.

In this paper, we move a step forward to account for the intrinsic randomness in relevance judgements and we frame them into a general stochastic approach where the judgement assigned to a document is a binomial random variable whose expectation p indicates the quantity of relevance assigned to that document.

We show how to apply the proposed framework to the definition of *random evaluation measures*, i.e. IR evaluation measures able to incorporate the inherent randomness in relevance judgements, and how this new approach not only eliminates the distinction between binary and multi-graded evaluation measures but also deals with incomplete information, by providing us with a single unifying vision which can be coherently applied to all the IR evaluation measures.

We apply our framework to two widely known measures, namely *Average Precision (AP)* and *Rank-Biased Precision (RBP)* [10], in order to show the generality of the proposed solution. We also conduct a systematic experimentation using TREC collections which shows that these new random evaluation measures are a coherent extension of their non-random counterparts and that they have many desirable properties in terms of robustness to incomplete information and sensitivity in discriminating among systems.

The paper is organized as follows: Section 2 discusses some related works; Section 3 introduces our stochastic framework; Section 4 reports the evaluation of the proposed approach; and, Section 5 draws some conclusions and outlooks possible future works.

2 Related Work

To the best of our knowledge, this paper represents one of the first attempts to explicitly model relevance judgements as a stochastic process with the specific goal to introduce random IR evaluation measures, benefiting from a single unifying view on multi-graded measures and incomplete information.

One of the closest areas is dealing with incomplete information in relevance judgments, i.e. how to account for unjudged documents [2, 11, 14, 18]. All these works differ from our approach in that they focus on unjudged documents in the pool and how to reliably estimate a proportion of relevant documents for them. On the contrary, we model each single relevance judgement as a binomial random variable and we derive a general stochastic framework where evaluation measures account for randomness in the assessment of each retrieved document, both judged and unjudged documents in the pool.

Finally, when it comes to multi-graded judgements, either we have evaluation measures which are natively multi-graded, such as *Normalized Discounted Cumulated Gain (nDCG)* [6] and *Expected Reciprocal Rank (ERR)* [3], or extensions from the binary to the multi-graded case, such as *Graded Average Precision (GAP)* [12]. However, all these cases treat relevance judgements as deterministic and the extensions from the binary to the multi-graded case are typically ad-hoc, i.e., they work only for a specific measure, while our approach is general and can be seamlessly applied to any IR evaluation measure.

3 Proposed Stochastic Model

3.1 Random Relevance

We stem from the notation proposed by [4] for defining the basic concepts of topics, documents, ground-truth, run and judged run and we extend it to account for random relevance instead of deterministic one.

Let us consider a set of **documents** D and a set of **topics** T. Let $(REL, \preceq)$ be a totally ordered set of **relevance degrees**, where we assume the existence of a minimum that we call the **non-relevant** relevance degree $\mathtt{nr} = \min(REL)$. We assume that REL is a finite set. Moreover, given $m \in \mathbb{N}$ such that $|REL| = m+1$, we denote its strictly ordered elements as $rel_0 \prec \cdots \prec rel_m$, where $rel_0 = \mathtt{nr}$.

For each pair $(t, d) \in T \times D$, the **ground-truth** GT is a map which assigns a relevance degree $rel \in REL$ to a document d with respect to a topic t. The **recall base** is the map RB from T into $\mathbb{N}$ defined as the total number of relevant documents for a given topic $t \mapsto RB_t = \big|\{d \in D : GT(t, d) \succ \mathtt{nr}\}\big|$.

Given a run $r_t = (d_1, \ldots, d_N)$ of length N, let $\hat{r}_t[i]$ be the relevance assigned to the document d_i for the topic t, i.e. $\hat{r}_t[i] = GT(t, d_i)$.

Given a positive integer N, the length of the run, we define the **set of retrieved documents** as $D(N) = \{(d_1, \ldots, d_N) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. the ranked list of retrieved documents without duplicates, and the **universe set of retrieved documents** as $\mathcal{D} := \bigcup_{N=1}^{|D|} D(N)$.

The already existing binary (when $m = 1$) and multi-graded ($m > 1$) evaluation measures usually map each relevance degree into an integer number. For example, if $REL = \{\mathtt{nr}, \mathtt{r}\}$, then AP assigns the value 0 to every non-relevant document while 1 is used for the relevant ones. Similarly, if $REL = \{\mathtt{nr}, \mathtt{pr}, \mathtt{r}, \mathtt{hr}\}$, nDCG [6] assigns an integer number to each relevance degree, e.g. 0, 5, 10 and 15, consistently with the ordering among the relevance degrees. If it is very natural to assign 0 to a non-relevant document and 1 to a relevant one, being this latter value just any possible positive number different from zero that simply indicates the "presence" of some relevance, the situation is not so clear in the case of multi-graded relevance. For example, if 5 is the value assigned to a partially relevant ($\mathtt{pr}$) document and 10 is the one for a relevant document ($\mathtt{r}$), this does not necessary mean that relevant documents are twice as relevant as partially relevant ones, even though their contribution to some measures, e.g. nDCG, is actually doubled.

Could there exist a right or at least a common way to assign integers to different degrees of relevance? For example, [9] proposed magnitude estimation as a way to let users to estimate relevance on their own scale and raised the question whether a single view of relevance is actually appropriate to describe a population of users. The answer to this question is not easy and in the present paper, to account for a population of users, we consider the relevance of each document as a **random** number chosen between $\{0, 1\}$, where again 0 means completely "non-relevant" and 1 means "fully relevant".

Therefore, we describe the relevance of a document via a **binomial random variable** $B(1, p)$ with parameters 1 and p, where p roughly defines the *quantity of relevance* of that document. Recall that such a binomial random variable is a function from Ω, i.e. a suitable sample space, into $\{0, 1\}$ and it is equal to 1 with probability p and 0 with probability $1 - p$.

In accordance with this construction, we redefine the ground-truth as follows: for each pair $(t, d_i) \in T \times D$, the **random Ground-truth** RGT, also called random relevance, is a binomial random variable of parameters $(1, p_{t,d_i})$, where

p_{t,d_i} is the parameter associated to the document d_i with respect to a topic t. $p_{t,d_i} = 0$ corresponds to a document completely not relevant and $p_{t,d_i} = 1$ to a fully relevant document. For simplicity, in the sequel we will write $p_{t,i}$ instead of p_{t,d_i}. Moreover, we replace the deterministic recall base RB_t defined before with $\widehat{RB}_t$, the expected total relevance present in $\mathcal{D}$, i.e. $\widehat{RB}_t = \sum_{d \in \mathcal{D}} \mathbb{E}\big[RGT(t,d)\big]$ whose true value will be most of the times just estimated.

Let $\mathcal{R}$ be the set $\bigcup_{N=1}^{|D|} \{0,1\}^N$; a **random judged run** is the function $\hat{r}_t$ from $\Omega \times T \times \mathcal{D}$ into $\mathcal{R}$, which assigns a random relevance to each retrieved document in the ranked list

$$(\omega, t, r_t) \mapsto \hat{r}_t(\omega) = \big(RGT(t,d_1)(\omega), \ldots, RGT(t,d_N)(\omega)\big) \ .$$

3.2 Random Evaluation Measures

Generally speaking, a *random evaluation measure* is an application

$$M : \Omega \times T \times D \to \mathbb{R}_+$$

obtained by the composition of the random judged run with the map

$$\mu : \mathcal{R} \to \mathbb{R}_+$$

giving $M = \mu\big(RGT(t,d_1)(\omega), \ldots, RGT(t,d_N)(\omega)\big)$.

To show how to apply the proposed approach, we provide the definition of the random version of two well-known evaluation measures, namely RBP and AP.

Random Rank Biased Precision (RRBP) of parameter $q \in (0,1)$ is defined as

$$RBP[\hat{r}_t(\omega)] = (1-q) \sum_{n=1}^{N} q^{n-1} \hat{r}_t[n](\omega) \ .$$

where q denotes the persistence of the user in scanning the results list.

Random Average Precision (RAP) is defined as

$$AP[\hat{r}_t(\omega)] = \frac{1}{\widehat{RB}_t} \sum_{n=1}^{N} \left(\frac{1}{n} \sum_{m=1}^{n} \hat{r}_t[m](\omega) \right) \hat{r}_t[n](\omega) \ .$$

To compare different systems, we need to define an ordering among runs of documents. Since the relevance is now stochastic, the ordering of the systems has to be defined in terms of the laws of the random relevances of the documents retrieved in the runs.

Definition 1. *Given a topic t, two runs of documents r_t and s_t and a random evaluation measure $M(\cdot,t)(\omega)$, we define a weak order on $\mathcal{R}$ as*

$$r_t \preceq s_t \quad \Leftrightarrow \quad \mathbb{E}[M(r_t,t)(\omega)] \leq \mathbb{E}[M(s_t,t)(\omega)] \ .$$

Therefore let us now take into account the expectations of the random versions of RBP and AP. We assume the random relevances of different documents to be independent random variables.

We define the **expected Rank Biased Precision (eRRBP)** as the expectation of RBP when computed over runs with random relevance degrees. Since RBP is a linear combination of independent random variables, the computation of its mean is quite simple, giving rise to the following expression:

$$\mathbb{E}\big[RBP[\hat{r}_t(\omega)]\big] = (1-q)\sum_{n=1}^{N} q^{n-1} p_{t,n} \ . \tag{1}$$

Similarly, **expected Random Average Precision (eRAP)** is the expectation of AP, whose computation is slightly more complicated, since we here have the sum of partial sums of the same random variables. The mean is:

$$\mathbb{E}\big[AP[\hat{r}_t(\omega)]\big] = \frac{1}{\widehat{RB}_t}\sum_{n=1}^{N}\frac{1}{n}\left(1+\sum_{m=1}^{n-1} p_{t,m}\right) p_{t,n} \ , \tag{2}$$

where we have made use of the fact that all the moments of a $B(1, p_{t,k})$ random variable are equal to $p_{t,k}$.

Summing up, the proposed random measures decouple the problem of determining the presence of relevance from that of indicating the amount of relevance. Indeed, the former is represented by the output of the binomial random variables, either 0 in case of absence of relevance or 1 in case of presence of relevance; the latter is represented instead by the parameter p of the binomial random variables, which accounts for the amount of relevance. In this way, the same mechanism for indicating the presence and amount of relevance is used for both the binary and multi-graded case, thus eliminating the distinction between them. Furthermore, these random measures allow us to "seed" some relevance also for the not relevant documents by setting the parameter p slightly greater than 0 in that case. This is especially useful in the case of unjudged documents and incomplete information, since it allows us to somehow capture what we might call the "dark relevance" present in the document's universe. For these reasons, we can say that the proposed random measures are able to seamlessly describe both multi-graded and incomplete information.

4 Experiments

We focus on the following existing evaluation measures to compare ours against: nDCG [6] and ERR [3] as examples of natively multi-graded evaluation measures; GAP [12] as an example of extension of AP to graded judgments; and *Graded Rank-Biased Precision (gRBP)* [15] as an example of use of RBP [10] with graded judgements; *Binary Preference (bpref)* [2] and *Inferred Average Precision (infAP)* [18] as examples of binary measures for incomplete information.

We used the following collections: *TREC Terabyte track* `T14` using the `GOV2` collection with 50 topics, deep pools at depth 100, and graded relevance judgments – i.e., not relevant, relevant and highly relevant; 58 runs were submitted, retrieving 1,000 documents for each topic; *TREC Web track* `T21` using the `ClueWeb09` collection with 50 topics, shallow pools using depths 20 and 30, and graded relevance judgments – i.e., junk, not relevant, relevant, highly relevant, key and nav; we considered junk and not relevant as a single not relevant level and key and nav as a single key level; 27 runs were submitted, retrieving 10,000 documents for each topic.

For nDCG we use a log base $b = 10$ and gains 0, 5, 10, and 15 for not relevant, relevant, highly relevant, and key documents, respectively. For ERR we instead use 0, 1, 2 and 3 as gains. For RBP we set the persistence parameter q to 0.8, which works well for both deep and shallow pools as pointed out by [10].

Although our approach provides a very fine-grained level of detail in defining the random relevance up to each (topic, document) pair, e.g. by using magnitude estimation techniques [9], in the following evaluation we let the parameter $p_{t,d}$ to be fixed for each relevance degree to a value p_k, independently from the document at hand, since this is the way in which all the IR measures we compare against work and this is the information available in the pools of the used collections. We can view each p_k as how much an assessor is confident that every given document with relevance degree equal to rel_k is actually relevant. The different values of the parameters p_k are reported in the caption of the figures which display the experimental results later on.

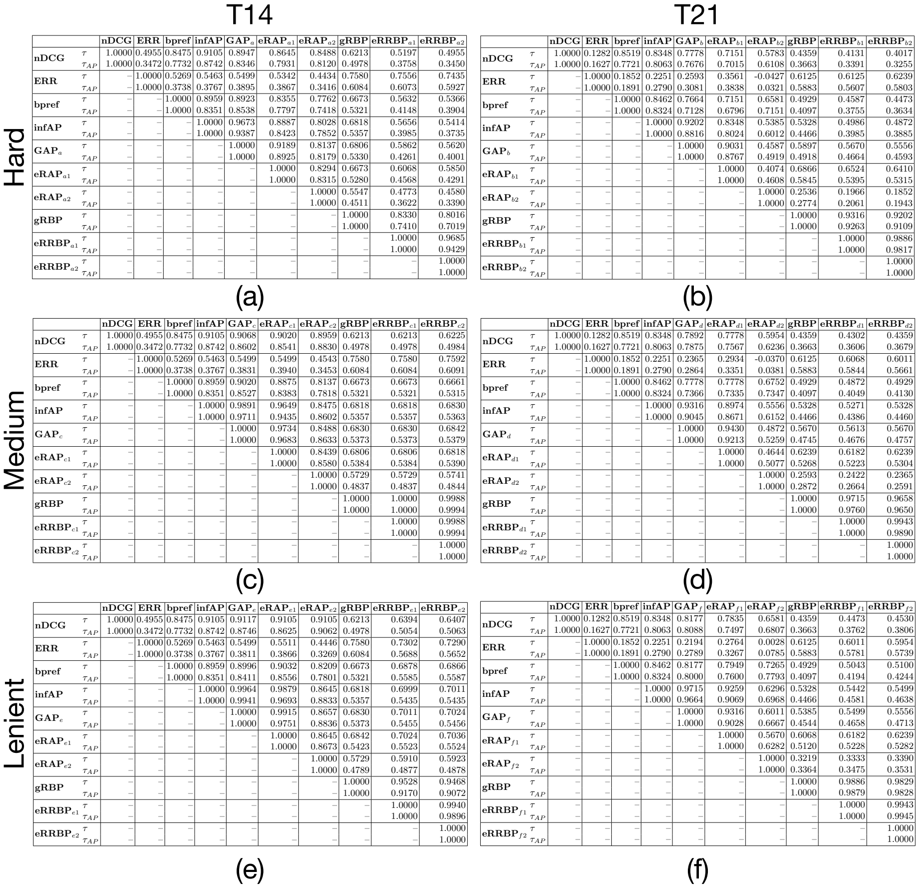
To ease the reproducibility of the experiments, the code for running them is available at: `https://bitbucket.org/frrncl/ecir2018`.

4.1 RQ1: Relation to Other Evaluation Measures

Figures 1 report the outcomes of the correlation analysis on `T14` and `T21`, respectively, using both Kendall's τ correlation [8] and τ_{ap} correlation [19]. Each row represents an alternative configuration of the parameters ranging from hard to lenient in the sense that, for example in the case of three relevance degrees, GAP with threshold probabilities $[0.00, 1.00]$ corresponds to AP when you perform a hard mapping to binary relevance, i.e. only the top relevance degree is considered relevant; on the other hand, GAP with threshold probabilities $[1.00, 0.00]$ corresponds to AP when you perform a lenient mapping to binary relevance, i.e. every relevance degree above not relevant is considered relevant.

For each set of parameters (hard, medium, lenient), we explore two options for eRAP and eRRBP. Option 1 makes eRAP to behave as close as possible to GAP by constraining the eRAP probabilities à la GAP: for example, in the case of three relevance degrees if the GAP threshold probabilities are $[g_1, g_2]$ we constraint the eRAP probabilities to $[0, g_1, g_1+g_2]$. Option 2 lets eRAP to behave in its intended way of use with more freedom in the choice of the probabilities, still being in the hard, medium or lenient cases.

Note that nDCG, ERR, and gRBP are always the same in all the three cases, i.e. hard, medium and lenient case, since they do not depend on different ways


T14
T21
Hard
Medium
Lenient

		nDCG	ERR	bpref	infAP	GAP$_a$	eRAP$_{a1}$	eRAP$_{a2}$	gRBP	eRRBP$_{a1}$	eRRBP$_{a2}$
nDCG	τ	1.0000	0.4955	0.8475	0.9105	0.8947	0.8645	0.8488	0.6213	0.5197	0.4955
	τ_{AP}	1.0000	0.3472	0.7732	0.8742	0.8346	0.7931	0.8120	0.4978	0.3758	0.3450
ERR	τ	–	1.0000	0.5269	0.5463	0.5499	0.5342	0.4434	0.7580	0.7556	0.7435
	τ_{AP}	–	1.0000	0.3738	0.3767	0.3895	0.3867	0.3416	0.6084	0.6073	0.5927
bpref	τ	–	–	1.0000	0.8959	0.8923	0.8355	0.7762	0.6673	0.5632	0.5366
	τ_{AP}	–	–	1.0000	0.8351	0.8538	0.7797	0.7418	0.5321	0.4148	0.3904
infAP	τ	–	–	–	1.0000	0.9673	0.8887	0.8028	0.6818	0.5656	0.5414
	τ_{AP}	–	–	–	1.0000	0.9387	0.8423	0.7852	0.5357	0.3985	0.3735
GAP$_a$	τ	–	–	–	–	1.0000	0.9189	0.8137	0.6806	0.5862	0.5620
	τ_{AP}	–	–	–	–	1.0000	0.8925	0.8179	0.5330	0.4261	0.4001
eRAP$_{a1}$	τ	–	–	–	–	–	1.0000	0.8294	0.6673	0.6068	0.5850
	τ_{AP}	–	–	–	–	–	1.0000	0.8315	0.5280	0.4568	0.4291
eRAP$_{a2}$	τ	–	–	–	–	–	–	1.0000	0.5547	0.4773	0.4580
	τ_{AP}	–	–	–	–	–	–	1.0000	0.4511	0.3622	0.3390
gRBP	τ	–	–	–	–	–	–	–	1.0000	0.8330	0.8016
	τ_{AP}	–	–	–	–	–	–	–	1.0000	0.7410	0.7019
eRRBP$_{a1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9685
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9429
eRRBP$_{a2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(a)

		nDCG	ERR	bpref	infAP	GAP$_b$	eRAP$_{b1}$	eRAP$_{b2}$	gRBP	eRRBP$_{b1}$	eRRBP$_{b2}$
nDCG	τ	1.0000	0.1282	0.8519	0.8348	0.7778	0.7151	0.5783	0.4359	0.4131	0.4017
	τ_{AP}	1.0000	0.1627	0.7721	0.8063	0.7676	0.7015	0.6108	0.3663	0.3391	0.3255
ERR	τ	–	1.0000	0.1852	0.2251	0.2593	0.3561	-0.0427	0.6125	0.6125	0.6239
	τ_{AP}	–	1.0000	0.1891	0.2790	0.3081	0.3838	0.0321	0.5883	0.5607	0.5803
bpref	τ	–	–	1.0000	0.8462	0.7664	0.7151	0.6581	0.4929	0.4587	0.4473
	τ_{AP}	–	–	1.0000	0.8324	0.7128	0.6796	0.7151	0.4097	0.3755	0.3634
infAP	τ	–	–	–	1.0000	0.9202	0.8348	0.5385	0.5328	0.4986	0.4872
	τ_{AP}	–	–	–	1.0000	0.8816	0.8024	0.6012	0.4466	0.3985	0.3885
GAP$_b$	τ	–	–	–	–	1.0000	0.9031	0.4587	0.5897	0.5670	0.5556
	τ_{AP}	–	–	–	–	1.0000	0.8767	0.4919	0.4918	0.4664	0.4593
eRAP$_{b1}$	τ	–	–	–	–	–	1.0000	0.4074	0.6866	0.6524	0.6410
	τ_{AP}	–	–	–	–	–	1.0000	0.4608	0.5845	0.5395	0.5315
eRAP$_{b2}$	τ	–	–	–	–	–	–	1.0000	0.2536	0.1966	0.1852
	τ_{AP}	–	–	–	–	–	–	1.0000	0.2774	0.2061	0.1943
gRBP	τ	–	–	–	–	–	–	–	1.0000	0.9316	0.9202
	τ_{AP}	–	–	–	–	–	–	–	1.0000	0.9263	0.9109
eRRBP$_{b1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9886
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9817
eRRBP$_{b2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(b)

		nDCG	ERR	bpref	infAP	GAP$_c$	eRAP$_{c1}$	eRAP$_{c2}$	gRBP	eRRBP$_{c1}$	eRRBP$_{c2}$
nDCG	τ	1.0000	0.4955	0.8475	0.9105	0.9068	0.9020	0.8959	0.6213	0.6213	0.6225
	τ_{AP}	1.0000	0.3472	0.7732	0.8742	0.8602	0.8541	0.8830	0.4978	0.4978	0.4984
ERR	τ	–	1.0000	0.5269	0.5463	0.5499	0.5499	0.4543	0.7580	0.7580	0.7592
	τ_{AP}	–	1.0000	0.3738	0.3767	0.3831	0.3940	0.3453	0.6084	0.6084	0.6091
bpref	τ	–	–	1.0000	0.8959	0.9020	0.8875	0.8137	0.6673	0.6673	0.6661
	τ_{AP}	–	–	1.0000	0.8351	0.8527	0.8383	0.7818	0.5321	0.5321	0.5315
infAP	τ	–	–	–	1.0000	0.9891	0.9649	0.8475	0.6818	0.6818	0.6830
	τ_{AP}	–	–	–	1.0000	0.9711	0.9435	0.8602	0.5357	0.5357	0.5363
GAP$_c$	τ	–	–	–	–	1.0000	0.9734	0.8488	0.6830	0.6830	0.6842
	τ_{AP}	–	–	–	–	1.0000	0.9683	0.8633	0.5373	0.5373	0.5379
eRAP$_{c1}$	τ	–	–	–	–	–	1.0000	0.8439	0.6806	0.6806	0.6818
	τ_{AP}	–	–	–	–	–	1.0000	0.8580	0.5384	0.5384	0.5390
eRAP$_{c2}$	τ	–	–	–	–	–	–	1.0000	0.5729	0.5729	0.5741
	τ_{AP}	–	–	–	–	–	–	1.0000	0.4837	0.4837	0.4844
gRBP	τ	–	–	–	–	–	–	–	1.0000	1.0000	0.9988
	τ_{AP}	–	–	–	–	–	–	–	1.0000	1.0000	0.9994
eRRBP$_{c1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9988
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9994
eRRBP$_{c2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(c)

		nDCG	ERR	bpref	infAP	GAP$_d$	eRAP$_{d1}$	eRAP$_{d2}$	gRBP	eRRBP$_{d1}$	eRRBP$_{d2}$
nDCG	τ	1.0000	0.1282	0.8519	0.8348	0.7892	0.7778	0.5954	0.4359	0.4302	0.4359
	τ_{AP}	1.0000	0.1627	0.7721	0.8063	0.7875	0.7567	0.6236	0.3663	0.3606	0.3679
ERR	τ	–	1.0000	0.1852	0.2251	0.2365	0.2934	-0.0370	0.6125	0.6068	0.6011
	τ_{AP}	–	1.0000	0.1891	0.2790	0.2864	0.3351	0.0381	0.5883	0.5844	0.5661
bpref	τ	–	–	1.0000	0.8462	0.7778	0.7778	0.6752	0.4929	0.4872	0.4929
	τ_{AP}	–	–	1.0000	0.8324	0.7366	0.7335	0.7347	0.4097	0.4049	0.4130
infAP	τ	–	–	–	1.0000	0.9316	0.8974	0.5556	0.5328	0.5271	0.5328
	τ_{AP}	–	–	–	1.0000	0.9045	0.8671	0.6152	0.4466	0.4386	0.4460
GAP$_d$	τ	–	–	–	–	1.0000	0.9430	0.4872	0.5670	0.5613	0.5670
	τ_{AP}	–	–	–	–	1.0000	0.9213	0.5259	0.4745	0.4676	0.4757
eRAP$_{d1}$	τ	–	–	–	–	–	1.0000	0.4644	0.6239	0.6182	0.6239
	τ_{AP}	–	–	–	–	–	1.0000	0.5077	0.5268	0.5223	0.5304
eRAP$_{d2}$	τ	–	–	–	–	–	–	1.0000	0.2593	0.2422	0.2365
	τ_{AP}	–	–	–	–	–	–	1.0000	0.2872	0.2664	0.2591
gRBP	τ	–	–	–	–	–	–	–	1.0000	0.9715	0.9658
	τ_{AP}	–	–	–	–	–	–	–	1.0000	0.9760	0.9650
eRRBP$_{d1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9943
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9890
eRRBP$_{d2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(d)

		nDCG	ERR	bpref	infAP	GAP$_e$	eRAP$_{e1}$	eRAP$_{e2}$	gRBP	eRRBP$_{e1}$	eRRBP$_{e2}$
nDCG	τ	1.0000	0.4955	0.8475	0.9105	0.9117	0.9105	0.9105	0.6213	0.6394	0.6407
	τ_{AP}	1.0000	0.3472	0.7732	0.8742	0.8746	0.8625	0.9062	0.4978	0.5054	0.5063
ERR	τ	–	1.0000	0.5269	0.5463	0.5499	0.5511	0.4446	0.7580	0.7302	0.7290
	τ_{AP}	–	1.0000	0.3738	0.3767	0.3811	0.3866	0.3269	0.6084	0.5688	0.5652
bpref	τ	–	–	1.0000	0.8959	0.8996	0.9032	0.8209	0.6673	0.6878	0.6866
	τ_{AP}	–	–	1.0000	0.8351	0.8411	0.8556	0.7801	0.5321	0.5585	0.5587
infAP	τ	–	–	–	1.0000	0.9964	0.9879	0.8645	0.6818	0.6999	0.7011
	τ_{AP}	–	–	–	1.0000	0.9941	0.9693	0.8833	0.5357	0.5435	0.5435
GAP$_e$	τ	–	–	–	–	1.0000	0.9915	0.8657	0.6830	0.7011	0.7024
	τ_{AP}	–	–	–	–	1.0000	0.9751	0.8836	0.5373	0.5455	0.5456
eRAP$_{e1}$	τ	–	–	–	–	–	1.0000	0.8645	0.6842	0.7024	0.7036
	τ_{AP}	–	–	–	–	–	1.0000	0.8673	0.5423	0.5523	0.5524
eRAP$_{e2}$	τ	–	–	–	–	–	–	1.0000	0.5729	0.5910	0.5923
	τ_{AP}	–	–	–	–	–	–	1.0000	0.4789	0.4877	0.4878
gRBP	τ	–	–	–	–	–	–	–	1.0000	0.9528	0.9468
	τ_{AP}	–	–	–	–	–	–	–	1.0000	0.9170	0.9072
eRRBP$_{e1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9940
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9896
eRRBP$_{e2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(e)

		nDCG	ERR	bpref	infAP	GAP$_f$	eRAP$_{f1}$	eRAP$_{f2}$	gRBP	eRRBP$_{f1}$	eRRBP$_{f2}$
nDCG	τ	1.0000	0.1282	0.8519	0.8348	0.8177	0.7835	0.6581	0.4359	0.4473	0.4530
	τ_{AP}	1.0000	0.1627	0.7721	0.8063	0.8088	0.7497	0.6807	0.3663	0.3762	0.3806
ERR	τ	–	1.0000	0.1852	0.2251	0.2194	0.2764	0.0028	0.6125	0.6011	0.5954
	τ_{AP}	–	1.0000	0.1891	0.2790	0.2789	0.3267	0.0785	0.5883	0.5781	0.5739
bpref	τ	–	–	1.0000	0.8462	0.8177	0.7949	0.7265	0.4929	0.5043	0.5100
	τ_{AP}	–	–	1.0000	0.8324	0.8000	0.7600	0.7793	0.4097	0.4194	0.4244
infAP	τ	–	–	–	1.0000	0.9715	0.9259	0.6296	0.5328	0.5442	0.5499
	τ_{AP}	–	–	–	1.0000	0.9664	0.9069	0.6968	0.4466	0.4581	0.4638
GAP$_f$	τ	–	–	–	–	1.0000	0.9316	0.6011	0.5385	0.5499	0.5556
	τ_{AP}	–	–	–	–	1.0000	0.9028	0.6667	0.4544	0.4658	0.4713
eRAP$_{f1}$	τ	–	–	–	–	–	1.0000	0.5670	0.6068	0.6182	0.6239
	τ_{AP}	–	–	–	–	–	1.0000	0.6282	0.5120	0.5228	0.5282
eRAP$_{f2}$	τ	–	–	–	–	–	–	1.0000	0.3219	0.3333	0.3390
	τ_{AP}	–	–	–	–	–	–	1.0000	0.3364	0.3475	0.3531
gRBP	τ	–	–	–	–	–	–	–	1.0000	0.9886	0.9829
	τ_{AP}	–	–	–	–	–	–	–	1.0000	0.9879	0.9828
eRRBP$_{f1}$	τ	–	–	–	–	–	–	–	–	1.0000	0.9943
	τ_{AP}	–	–	–	–	–	–	–	–	1.0000	0.9945
eRRBP$_{f2}$	τ	–	–	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	–	–	1.0000

(f)

Fig. 1. Correlation analysis on `T14` (first column) and `T21` (second column) for different sets of parameters. The first row of subfigures reports the *hard case*: (a) GAP with threshold probabilities $a = [0.20, 0.80]$, eRAP end eRRBP with probabilities $a1 = [0.00, 0.20, 1.00]$ and $a2 = [0.05, 0.20, 0.95]$ on `T14`; (b) GAP with threshold probabilities $b = [0.10, 0.40, 0.50]$, eRAP end eRRBP with probabilities $b1 = [0.00, 0.10, 0.50, 1.00]$ and $b2 = [0.05, 0.10, 0.50, 0.95]$ on `T21`. The second row of subfigures reports the *medium case*: (c) GAP with threshold probabilities $c = [0.50, 0.50]$, eRAP end eRRBP with probabilities $c1 = [0.00, 0.50, 1.00]$ and $c2 = [0.05, 0.50, 0.95]$ on `T14`; (d) GAP with threshold probabilities $d = [0.20, 0.40, 0.40]$, eRAP end eRRBP with probabilities $d1 = [0.00, 0.20, 0.60, 1.00]$ and $d2 = [0.05, 0.20, 0.60, 0.95]$ on `T21`. The third row of subfigures reports the *lenient case*: (e) GAP with threshold probabilities $e = [0.70, 0.30]$, eRAP end eRRBP with probabilities $e1 = [0.00, 0.70, 1.00]$ and $e2 = [0.05, 0.70, 0.95]$ on `T14`; (f) GAP with threshold probabilities $f = [0.40, 0.40, 0.20]$, eRAP end eRRBP with probabilities $f1 = [0.00, 0.40, 0.80, 1.00]$ and $f2 = C2 = [0.05, 0.40, 0.80, 0.95]$ on `T21`.

of thresholding the relevance degree. The same holds for bpref and infAP which always adopt a lenient mapping to binary relevance.

We can observe a very general trend on both T14 and T21: in the case of GAP and eRAP moving from the hard case to the lenient case increases the correlation with nDCG in terms of both τ and τ_{ap}, indicating that somehow "seeing" more relevance degrees makes these evaluation measures closer to a natively multi-graded one. However, while in the case of T14 and three relevance degrees these correlations are quite high, around or above 0.9 in terms of Kendall's τ, when it comes to T21 and four relevance degrees they are typically below 0.8 in terms of Kendall's τ; this suggest that, as the number of relevance degrees increases, these measures tend to take a different angle on what multi-graded relevance is.

We can observe a similar behaviour also for eRRBP with respect to nDCG, while the correlation between gRBP and nDCG is the same in all the cases since they do not depend on the choice of the probabilities; we can also see how the correlation between nDCG and eRRBP is lower than the one between nDCG and gRBP in the hard case, somehow similar in the medium case, and higher in the lenient case. However, these correlations are generally low, around or below 0.60 for T14 and 0.45 for T21 in terms of Kendall's τ, suggesting an approach to multi-graded relevance quite different from nDCG.

When it comes to ERR we can see a similar increasing correlation trend with GAP and eRAP, even if the correlations are very low – below 0.55 for T14 and 0.35 for T21 in terms of Kendall's τ – denoting completely different approaches to ranking systems, probably due to the extremely top-heavy nature of ERR. On the other hand, the correlation between ERR and gRBP/eRRBP is much higher – around 0.75 for T14 and 0.60 for T21 in terms of Kendall's τ – suggesting a greater agreement probably due to the more top-heavy nature of gRBP/eRRBP.

When it comes to the comparison between GAP and eRAP we can note that eRAP with probabilities constrained à la GAP (option 1) is very highly correlated with GAP with both τ and τ_{ap} almost always well above 0.9 on both T14 and T21; moreover, this correlation tends to increase moving from the hard to the lenient cases.

If we consider GAP and eRAP used in its intended way (option 2), we can observe high correlations in the case of T14 above 0.8 for both τ and τ_{ap} while they drop below 0.65 in the case of T21, indicating that the more the number of relevance degrees the more GAP and eRAP depart from each other; moreover, as in the previous cases, we can note an increasing trend as we pass from the hard to the lenient cases.

The comparison between gRBP and eRRBP shows that there is not much difference between using option 1 and 2, since both are very highly correlated with gRBP, being τ and τ_{ap} above 0.9 on both T14 and T21. This hints that the intrinsic structure of the RBP somehow "prevails" on the way in which you make it multi-graded and, as a result, all the different ways to make it multi-graded turn out to be very correlated.

When it comes to the comparison with measures for incomplete information, i.e., bpref and infAP, we can observe that there is an increasing correlation

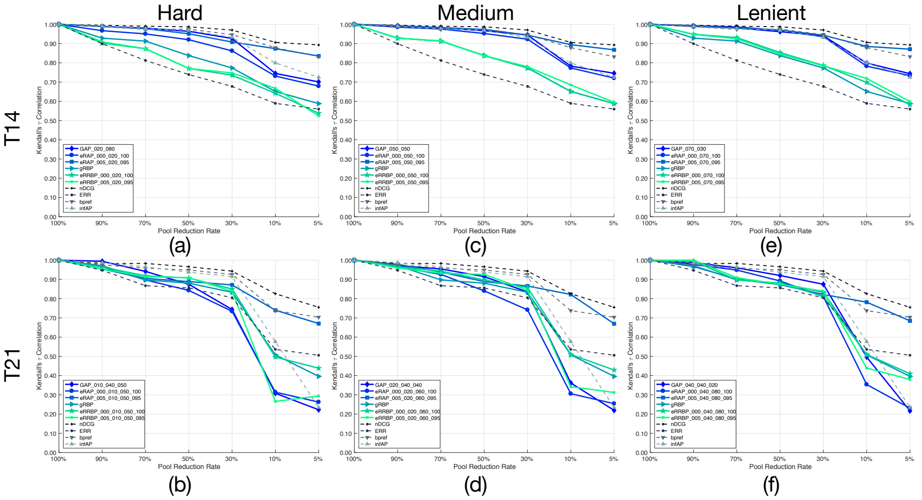


Fig. 2. Robustness to pool downsampling for both `T14` (first row) and `T21` (second row). The probabilities for the hard, medium, and lenient cases are the same as in Figure 1 for `T14` and `T21`.

trend passing from the hard to the lenient case for both eRAP and eRRBP on both `T14` and `T21`; this makes sense since both bpref and infAP adopt a lenient approach for mapping multi-graded judgements to binary ones. In particular, in the case of infAP and eRAP, we can observe quite high τ correlations, from 0.88 onwards on `T14` and from 0.83 onwards on `T21` when eRAP is constrained à la GAP (option 1) where slightly lower correlations on `T21` are due to the more multi-graded nature of this track. When we allow for more degrees of freedom in eRAP (option 2), correlations get lower, in the range 0.80-0.86 on `T14` and 0.53-0.63 on `T21`, still being coherent with infAP, but more affected by multi-graded judgments.

Overall, the correlation analysis shows how introducing the idea of random relevance and turning evaluation measures into random evaluation measures allows us to seamlessly manage both binary and multi-graded judgements, keeping a coherent vision with respect to both binary and multi-graded measures. Moreover, the same approach provide us also with an unifying view with respect to addressing incomplete information, as also investigated in the next section.

4.2 RQ2: Properties of the Evaluation Measures

Figure 2 shows the robustness of the evaluation measures to pool downsampling for both `T14` and `T21` in the hard, medium, and lenient cases considered before. We downsampled pools as in [2] and we computed the Kendall's τ correlation of each measure with respect to its version on the full pool as an indicator of how robust a measure is to pool downsampling. Consider that, downsampling

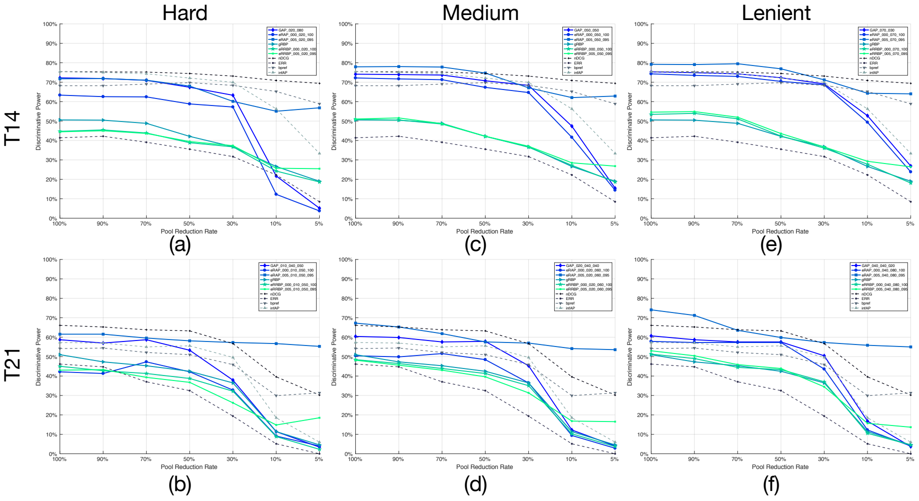


Fig. 3. Discriminative power at pool samples for both T14 (first row) and T21 (second row). The probabilities for the hard, medium, and lenient cases are the same as in Figure 1 for T14 and T21.

the pools, in the case of T14 we are passing from a deep to a shallow pool while for T21 we are passing from a shallow to an extremely shallow pool. We can see that, consistently with previous findings in the literature, nDCG, bpref, and infAP are among the most robust measures to pool downsampling while ERR is more sensitive to it.

In the case of T14 (first row of Figure 2) we can observe two separate clusters for all the cases (hard, medium, lenient): one for GAP and eRAP and the other for gRBP and eRRBP where the former is more robust to pool downsampling than the latter. In the case of GAP/eRAP we can see that the more we pass from the hard to the lenient case the closer is their behaviour to the nDCG, bpref, and infAP ones even if GAP has stable performances up to 30% reduction rate while eRAP used in its intended way (option 2, continuous blue line with squares) outperforms it and remains stable also for the more extreme reduction rates, basically behaving like nDCG and bpref.

In the case of T21 (second row of Figure 2), we do not have anymore two separate clusters but all the measures look very similar up to the 30% reduction rate, being nDCG and bpref always the top performing ones. However, for the 10% and 5% reduction rates all the measures have a sudden drop, even below the level of ERR, with the sole exceptions of nDCG, bpref, and eRAP used in its intended way (option 2, continuous blue line with squares) which remain stable and perform very similarly.

Figure 3 shows the *Discriminative Power (DP)* [13] achieved at different pool samples for both T14 and T21 in the hard, medium, and lenient cases.

If we consider the full pools, we can see that, consistently with previous findings in the literature, nDCG (DP = 75.44% on T14 and DP = 66.10% on T21) together with bpref (DP = 68.18% on T14 and DP = 54.13% on T21) and infAP (DP = 75.38% on T14 and DP = 57.26% on T21) are among the most discriminative measures while ERR (DP = 41.38% on T14 and DP = 46.15% on T21) is one of the least discriminative ones, due to its strongly top-heavy nature. As a general trend, we can see how the DP improves passing from the hard to the lenient case for both eRRBP and eRAP. In particular, we can see how gRBP (DP = 50.57% on T14 and DP = 51.00% on T21) and eRRBP (DP = 44.00%–55.00% on T14 and DP = 43.00%–53.00% on T21), due to their more top-heavy nature, behave somehow similarly to ERR and tend to be less discriminative than the other measures. On the other hand, GAP (DP = 72.00%–75.00% on T14 and DP = 58.00%–60.00% on T21) and eRAP (DP = 63.00%–79.00% on T14 and DP = 42.00%–74.00% on T21) behave closer to nDCG, bpref and infAP. In particular, on both T14 and T21, eRAP used in its intended way (option 2, continuous blue line with squares) outperforms even nDCG, bpref, and infAP in both the medium case (DP = 77.92% on T14 and DP = 67.24% on T21) and in the lenient case (DP = 79.19% on T14 and DP = 74.07% on T21).

If we consider the different down-sampled pools, we can observe that the above mentioned trends are roughly respected and the discriminative power is stable enough up to the 50% reduction rate. On the other hand, for higher pool reduction rates, there is typically a drop in the performance with the exception of nDCG, bpref, and eRAP used in its intended way (option 2, continuous blue line with squares). These three latter measures behave quite similarly on T14 while eRAP used in its intended way (option 2) substantially outperforms both nDCG and bpref on T21 in the hard, medium, and lenient cases.

Overall, this analysis suggests that the proposed random measures maintain (or even improve) desirable properties in terms of robustness to incomplete information and discriminative power but providing a single unified vision which account for binary and multi-graded judgements as well as for incomplete information.

5 Conclusions and Future Work

In this paper, we have proposed a general stochastic approach for modelling relevance as a random binomial variable. Besides modelling the intrinsic randomness present in the relevance assessment process and the different viewpoints of a user population, the random relevance allows us to turn evaluation measures into random evaluation measures and to provide a single unifying vision between binary and multi-graded measures as well as on the robustness to incomplete information.

A systematic experimentation on TREC collections has shown how these new random measures relate and differ from existing measures and that they have desirable properties in term of robustness to pool downsampling and ability of discriminating among different systems. Overall, this suggested that the proposed new stochastic approach provides the aforementioned benefit at no cost or even improving the properties of the random measures derived from it.

Future work will concern the investigation of further applications of the random relevance and the random evaluation measures. In the crowd-sourcing context, random relevance could be exploited as an alternative way, e.g. to majority voting, to merge pools produced by multiple crowd-assessors, since the relevance of a document could be expressed via a binomial random variable of parameter $(1, p)$, where p is determined accordingly to the different assessments given by the assessors.

References

1. Alonso, O., Mizzaro, S.: Using Crowdsourcing for TREC Relevance Assessment. IPM 48(6), 1053–1066 (2012)
2. Buckley, C., Voorhees, E.M.: Retrieval Evaluation with Incomplete Information. SIGIR 2004. pp. 25–32. (2004)
3. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected Reciprocal Rank for Graded Relevance. CIKM 2009. pp. 621–630. (2009)
4. Ferrante, M., Ferro, N., Maistro, M.: Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. ICTIR 2015. pp. 21–30. (2015)
5. Hosseini, M., Cox, I.J., Milić-Frayling, N., Kazai, G., Vinay, V.: On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. ECIR 2012. pp. 182–194. (2012)
6. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. TOIS 20(4), 422–446 2002)
7. Kazai, G., Craswell, N., Yilmaz, E., Tahaghoghi, S.S.M.: An Analysis of Systematic Judging Errors in Information Retrieval. CIKM 2012. pp. 105–114. (2012)
8. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
9. Maddalena, E., Mizzaro, S., Scholer, F., Turpin, A.: On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. TOIS 35(3), 19:1–19:32 (2017)
10. Moffat, A., Zobel, J.: Rank-biased Precision for Measurement of Retrieval Effectiveness. TOIS 27(1), 2:1–2:27 (2008)
11. Park, L.A.F.: Uncertainty in Rank-Biased Precision. ADCS 2016. pp. 73–76. (2016)
12. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending Average Precision to Graded Relevance Judgments. In: SIGIR 2010. pp. 603–610. (2010)
13. Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap. SIGIR 2006. pp. 525–532. (2006)
14. Sakai, T.: Alternatives to Bpref. SIGIR 2007. pp. 71–78. (2007)
15. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11(5), 447–470 (2008)
16. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. SIGIR 1998. pp. 315–323. (1998)
17. Webber, W., Chandar, P., Carterette, B.A.: Alternative Assessor Disagreement and Retrieval Depth. CIKM 2012. pp. 125–134. (2012)
18. Yilmaz, E., Aslam, J.A.: Estimating Average Precision With Incomplete and Imperfect Judgments. CIKM 2006. pp. 102–111. (2006)
19. Yilmaz, E., Aslam, J.A., Robertson, S.E.: A New Rank Correlation Coefficient for Information Retrieval. SIGIR 2008. pp. 587–594. (2008)
20. Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments. SIGIR 1998. pp. 307–314. (1998)