Exploiting Stopping Time to Evaluate Accumulated Relevance

Marco Ferrante
Department of Mathematics "Tullio Levi-Civita",
University of Padua, Italy
ferrante@math.unipd.it

Nicola Ferro
Department of Information Engineering,
University of Padua, Italy
ferro@dei.unipd.it

ABSTRACT

Evaluation measures are more or less explicitly based on user models which abstract how users interact with a ranked result list and how they accumulate utility from it. However, traditional measures typically come with a hard-coded user model which can be, at best, parametrized. Moreover, they take a deterministic approach which leads to assign a precise score to a system run.

In this paper, we take a different angle and, by relying on Markov chains and random walks, we propose a new family of evaluation measures which are able to accommodate for different and flexible user models, allow for simulating the interaction of different users, and turn the score into a random variable which more richly describes the performance of a system. We also show how the proposed framework allows for instantiating and better explaining some state-of-the-art measures, like AP, RBP, DCG, and ERR.

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**.

KEYWORDS

evaluation measure; Markov chain; stopping time; user model

ACM Reference Format:
Marco Ferrante and Nicola Ferro. 2020. Exploiting Stopping Time to Evaluate Accumulated Relevance. In *The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3409256.3409832

1 INTRODUCTION

System-oriented evaluation [25] abstracts away many details of how and why users interact with *Information Retrieval (IR)* systems in real settings, in order to provide a very controlled environment which allows for repeatedly running experiments in a replicable way. In this context, evaluation measures not only quantify the effectiveness of IR systems but they also bring back some notion of user by embedding the so-called *user models*, which provide an abridged template of the user behaviour when scanning and interacting with the ranked result list. Therefore, it becomes crucial how much realistic such user models are, since they shape how close to actual users we are in quantifying IR system performance.

Carterette [5] pointed out that "*model-based measures are actually composed from three distinct underlying models: (1) a* browsing model *that describes how a user interacts with results; (2) a model of* document utility*, describing how a user derives utility from individual relevant documents; (3) a* utility accumulation *model that describes how a user accumulates utility in the course of browsing.*" Like most papers in IR, Carterette considers as browsing model "*that of a user scanning down ranked results one-by-one and stopping at some rank k*" and, therefore, he models just the probability distribution of the *stopping rank*. As a model of document utility, he makes use of binary or graded relevance judgments and he presents four utility accumulation models which correspond to well-known evaluation measure, e.g. *Average Precision (AP)* [4] . While Carterette's approach does a remarkably good work in retro-fitting a single coherent framework on existing popular evaluation measures, in helping us to better understand their constituents, and in paving the road for a consistent definition of new evaluation measures, it still suffers from some limitations.

Sakai and Dou [24] pointed out that users do not move just forward while scanning a ranked result list of documents, but pretty often they move both forward and backward. As a matter of fact, query logs from Yandex [26], containing more than 30 million records, show that more than 20% of the sessions contain also backward transitions. This observation dramatically changes the way of modelling users. Indeed, it is no more sufficient to determine **how many** documents the user examined in a linear scan in order to be able to account for both the utility gained and the effort performed by the user. Instead, we need to consider **how much** the user moved (backward and forward) in the ranked result list before stopping her search. Moreover, also the notion of **which documents** have been visited, i.e. what the user gains, should be modified into **how many times** a document has been visited since, if the user visits the same document more than once, the derived utility may change at each visit. Moreover, the browsing models of Carterette are a sort of **macro scale description** of the user model, where we only define the probability distribution of the stopping rank and we abstract away a whole class of users, i.e. those with that stopping rank distribution, without describing how (each of) these users actually interact with the ranked result list. In addition, such browsing models are often very simple, i.e. a sequential scan, and somehow **hard-coded**, lacking the possibility of seamlessly specifying alternative browsing models for the same measure. On the other hand, the utility accumulation models may become complicated to accommodate a discount which depends on the visited rank positions which, conceptually, would better fit better as part of the browsing model instead. Finally, even within Carterette's approach, evaluation measures typically are real valued

ICTIR '20, September 14–17, 2020, Virtual Event, Norway

ACM ISBN 978-1-4503-8067-6/20/09…$15.00
https://doi.org/10.1145/3409256.3409832

deterministic functions and, at best, they can be parametrised. In this respect, Carterette et al. [6] observe that distributions are preferable to point estimates to better grasp user variability.

We overcome these limitations by changing what the browsing and document utility models are while keeping the utility accumulation model as simple and intuitive as the model of precision is, i.e. total accumulated utility – how many relevant documents have been retrieved in the case of precision – discounted by the effort performed by the user – the length of the run in the case of precision. We call **P@H** our new family of stochastic evaluation measures, being them inspired by the simplicity of precision. We define the browsing model as a stochastic process, in particular a Markovian process. Thus, we move to a **micro scale description**, where we still aggregate the behaviour of many users but instead we consider their individual dynamics. As a consequence, we do not hard code a specific user model into an evaluation measure but rather we take a **generative approach**, seamlessly specifying alternative user models in a single and coherent way. The browsing model incorporates the rule to determine the *stopping rank* H, that is in general a **stopping time** in the sense of the theory of the stochastic processes. In this way, the stopping rank depends on the dynamic chosen to describe the users, how they move forward and backward in the ranked result lists, and may be a finite as well as an infinite random variable.

The benefit of such a general approach is twofold. Firstly, we can define a measure of retrieval effectiveness as a function of the explicitly defined browsing model and its stopping rule, providing us both with **how many** documents the user examined and with **how much** she moved in the list. Secondly, these measures will be no more deterministic functions that map runs and topics into non negative numbers but **stochastic processes** that map runs and topics into **non negative random variables**. Moreover, this construction can be easily extended to more general Markovian dynamics and/or to a continuous-time setting, by defining the browsing model as a continuous-time stochastic process. We combine this new browsing model with a new document utility model accounting for the fact that a user can visit several times the same document. In this way, not only we comprise Carterette's notion of static and dynamic measures, i.e. independent of relevance judgement or dependent on a specific ranking, but we also generalise it by considering the utility gained by a user as dependent on **how many times** a document has been visited. Finally, being stochastic processes, this new generation of evaluation measures opens up a whole range of new possibilities when it comes to comparing and **ordering systems**. The most basic way of ordering systems is based on the expected value of such stochastic processes and this approach corresponds to the traditional way of ordering systems when you adopt an evaluation measure like AP. However, it is possible to envision also more sophisticated (partial or total) orderings, for example, based on the notion of *stochastic dominance* [31] and frame these new measures into the field of the probabilistic approach to measurement scales [23].

The paper is organised as follows: Section 2 presents the related works; Section 3 introduces our Markovian browsing model; Section 4 describes our document utility model; and, Section 5 defines our P@H measure and how to order runs. Section 6 shows how "classical" evaluation measures can be derived within our framework and how our approach can provide new insights about them. Section 7 draws some conclusions and outlooks for future work.

2 RELATED WORKS

Cooper [8] proposed *Expected Search Length (ESL)* where documents are weakly ordered in levels and users may move randomly within each level; he computed the expected number of steps needed to gather the desired number of relevant documents. We share with these authors the vision of a probabilistic random walk over results and the stochastic accumulation of relevance.

Robertson [22] introduced a probabilistic model based on the notion of the probability $p_s(n)$ of a user satisfaction point at rank n; AP and *Reciprocal Rank (RR)* are then derived by using different distributions $p_s(n)$ and by cumulating precision at different rank positions weighted by $p_s(n)$. A related probabilistic interpretation of AP has been proposed by Yilmaz and Aslam [32]. We adopt a congruent vision in that we propose a probabilistic accumulation model but driven by a explicit user dynamics and how many times a document has been visited.

Rank-Biased Precision (RBP) by Moffat and Zobel [20] adopts a simple user model where users sequentially scan the ranked result list with a probability p of advancing to the next rank position and a probability $1-p$ to stop; *Expected Reciprocal Rank (ERR)* by Chapelle et al. [7] adopts a cascading model where the probability of continuing search depends on the relevance of the visited documents; *Expected Browsing Utility (EBU)* by Yilmaz et al. [33] adopts a sequential scan approach as well but it distinguishes between clicked documents and relevant documents to model the stopping behaviour. Even if they rely on explicit user models, neither these previous measures are based on a Markovian approach nor they allow for backward and forward transitions.

Smucker and Clarke [27, 28] proposed *Time-Biased Gain (TBG)*, which explicitly considers the time spent in scanning documents, and they view it just as "*a semi-Markov model*"; note also that TBG assumes a sequential scanning of the ranked result list. Sakai and Dou [24] introduced the *U-measure* which allows for both forward and backward transitions but it does not adopt a Markovian approach. Ferrante et al. [11] defined *Markov Precision (MP)*, where they used the stationary distribution of a Markov chain to weight precision values by the (long-run) probability of a user actually visiting that rank position; in our case, we are not considering the stationary distribution of the Markov chain but rather the stopping time. Recently, van Dijk et al. [30] used Markov chains to model session-based IR but, differently from our case, they focused on specific layouts of the Markov chain specifically aimed at modelling the transition from one query to another one.

Recent developments led to the introduction of adaptive/dynamic user models. In the wake of ESL, Bailey et al. [2] introduced INST, a measure based on T – the number of relevant documents expected by a user – where the probability of users stopping the search increases as they get closer and closer to T. Zhang et al. [34] proposed the *Bejeweled Player Model (BPM)* which is characterized by upper limits for both benefits and costs, i.e. a user stops when she either has found sufficient useful information or has no more patience to continue. Azzopardi et al. [1] proposed a stopping model based on

the information foraging theory where users stop on the basis of a goal sensitive constraint, i.e. how much they expect to gain, and a rate sensitive constraint, i.e. until the rate of gain is high enough. We can embrace such adaptive user models within our framework by defining the transition probabilities depending on what the user has gained, how fast the user has gained, the effort needed to gain; the investigation on how to actually represent these adaptive user models within our framework will be focus of our future work.

Markovian approaches find an interesting application also when it comes to defining user-side and interactive IR models of search [19]. Baskaya et al. [3] proposed a searcher model where the main actions and phases (formulate query, scan a snippet, click a link, read a document, judge document relevant, stop session) are represented as states of a Markov chain together with the probabilities of transitioning from one state to another. Thomas et al. [29] and Maxwell [19] have extended the early model by Baskaya et al. [3] introducing more phases and a finer-grained description of transition among states. All these models could be somehow put in correspondence with the browsing model of Carterette [5], even if they do not have the specific goal defining an evaluation measure. Dungs and Fuhr [9] adopted *Hidden Markov Models (HMMs)* to distinguish between different search phases and to recognise them. Both Maxwell and Azzopardi [18] and Zhang et al. [35] relied on such kind of models to simulate user interaction and, eventually, evaluate IR systems. Our work is focused on how Markov chains can be exploited to better express user models from a system-oriented evaluation perspective. How to extend P@H to the case of user-oriented evaluation or even to the more complex case of *Interactive Information Retrieval (IIR)* is out-of-scope for this paper and it is left for future work.

Among others, Fuhr [15] raised the issue of the scales adopted by evaluation measures and how they affect the validity of the computed statistics, like mean and variance. Ferrante et al. [12, 13, 14] have developed a theory of IR evaluation measures and demonstrated the scale properties of several state-of-the-art measures. Recently, Ferrante et al. [10] have started to experimentally show the impact of departing from the scale assumptions behind evaluation measure. As discussed in the following, our framework can be used to express several state-of-the-art evaluation measures: some of them, such as precision and RBP with $p = 0.5$ are interval scales, some of them, such as AP or ERR, are not. Therefore, measures instantiated via P@H can be (or not) an interval scale. It is beyond the scope of the present work to investigate which constraints we need to impose on P@H in order to always ensure it generates interval-scale measures. Indeed, our purpose here is mainly to define a new family of stochastic measures and to show how we can leverage their distributional properties to order runs. Note, however, that the key issue raised by Ferrante et al. to determine the scale properties of an evaluation measure is the notion of order among runs. Therefore, the stochastic orders brought in by P@H offer an interesting opportunity of future investigation in this respect.

3 BROWSING MODEL

The browsing model is determined by a stochastic process $X = \{X_n, n \in \mathbb{N}\}$, where the random variable X_n denotes the rank position of the n-th document visited by the user. Therefore $\{X_n = i\}$ represents the event "the n-th document considered by the user is d_i". We assume that the stochastic process defined by the sequence of random variables $X_1, X_2, X_3, \ldots$ forms a Markov Chain.

Let us consider a run consisting of a ranked result list of N documents retrieved in response to a given topic t, where N can be a finite non-negative integer or $+\infty$. Let d_i be the document retrieved at rank position i. The ranked result list of documents is denoted with $\mathcal{D} = \{d_i, i \leq N\}$ and their ranks with $\mathcal{N} = \{1, 2, \ldots, N\}$.

The three basic ingredients needed to define a Markov chain are: (1) the state space S, i.e. the set of values of the random variables X_n; (2) the initial distribution, i.e. the distribution of the random variable X_1; and, (3) the transition probabilities matrix, i.e. the matrix whose entries are defined by $p_{i,j} = \mathbb{P}[X_{n+1} = j | X_n = i]$.

In the following, we assume that:

(1) The state space of the Markov chain is $S = \{End\} \cup \mathcal{N}$, where *End* stands for "end of the search" and $\mathcal{N}$ denotes the ranks of the retrieved documents.
(2) The initial distribution is $\mathbb{P}[X_1 = 1] = 1$, i.e. with probability 1 the user starts from the first document in the ranked result list, as assumed by most evaluation measures [7, 20, 28, 33].
(3) The transition probabilities satisfy the time homogeneous Markov property

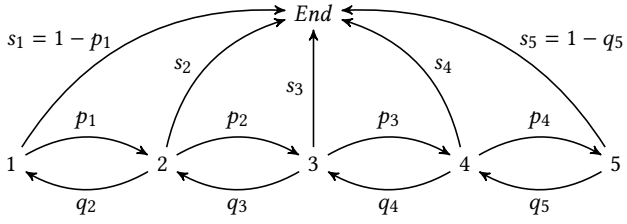
$$\mathbb{P}[X_{n+1} = j | X_n = i, \ldots, X_1 = i_1] = \mathbb{P}[X_{n+1} = j | X_n = i] = p_{i,j}$$

for any $n \in \mathbb{N}$ and $i, j, i_1, \ldots, i_{n-1} \in S$.
Moreover, we assume that $p_{End,End} = 1$, i.e. *End* is an absorbing state, and that $p_{i,j} = 0$ for any $i, j \in \mathcal{N}$ with $|i - j| \geq 2$, i.e. the user can move only between adjacent rank positions.

When $N < \infty$, we can represent the transition probabilities with the following stochastic matrix, where we denote the probabilities to step forward with p_i, to step backward with q_i and to stop during the search with s_i:

$$\begin{array}{c} \\ End \\ 1 \\ 2 \\ \vdots \\ N-2 \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{ccccccc} End & 1 & 2 & \ldots & N-2 & N-1 & N \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1-p_1 & 0 & p_1 & \cdots & 0 & 0 & 0 \\ s_2 & q_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ s_{N-2} & 0 & 0 & \cdots & 0 & p_{N-2} & 0 \\ s_{N-1} & 0 & 0 & \cdots & q_{N-1} & 0 & p_{N-1} \\ 1-q_N & 0 & 0 & \cdots & 0 & q_N & 0 \end{pmatrix} \end{array}$$

Example 3.1. *The following graph represents the browsing model in the case $N = 5$.*



Note that, due to the fact that any row of the transition matrix is a probability distribution, we have $p_1 + s_1 = 1, q_5 + s_5 = 1, p_i + q_i + s_i = 1$ for $i = 2, 3, 4$.

We can define three major classes of browsing models:

Deterministic Forward Browsing Model (DFBM) $p_i \equiv 1$, $q_i = s_i = 0$ for any $i < N$ and $q_N = 0$. This model corresponds to a user who sequentially scans each document in the ranked result list up to the last one.
As we will show in Section 6, this is the model behind precision; note that if you choose an N which is less than the length of a run, e.g. $N = 5$, you can obtain precision at the desired *Document Cut–off Value (DCV)*.
Stochastic Forward Browsing Model (SFBM) $p_i < 1$ for at least one $i < N$ and $q_i = 0$ for any i. This model corresponds to a user who, at document d_i, goes to the next document with probability p_i or stops her search with probability $1 - p_i$.
This model clearly resembles the model adopted by RBP but, as we will show in Section 6, it can be used also for other evaluation measures, such as AP, *Discounted Cumulated Gain (DCG)*, and ERR.
Random Walk Browsing Model (RWBM) $\max\{p_i, q_i\} < 1$ for at least one $i < N$. This corresponds to a user who, at document d_i, goes to the next document with probability p_i, or returns to the previous document with probability q_i, or stops her search with probability $s_i = 1 - p_i - q_i$.
This model is new to our framework and, to the best of our knowledge, no "classical" evaluation measure adopts it.

The values of p_i, q_i and s_i may depend on the rank i and on the relevance of the document at that rank and this allows us to accommodate for both static and dynamic measures as defined by Carterette [5].

We define the following random variable which represents the **stopping time** H of the search (see [21] for more details), i.e. the number of steps done by a single user before ending her search:

$$H^{End} = \inf\{n \geq 1 : X_n = End\} \,. \tag{1}$$

To simplify the exposition, we define $H = H^{End} - 1$, so X_N is the rank of the last document visited by the user.

Note that in the *deterministic forward browsing model* we have $H \equiv N$ while in the other two browsing models H is a non constant random variable with values on $\mathcal{N}$. Furthermore, if the number of documents is finite, in the case of the *stochastic forward browsing model* H is finite while in the case of the *random walk browsing model* it is an integer valued random variable.

4 DOCUMENT UTILITY MODEL

Since our browsing models allow for several visits to the same rank position, we have to take into account this and assume that at each visit the user may derive just part of the utility of a document.

If d_i denotes the document retrieved at rank position i and if t is a topic, we define $r_t[d_i] \in REL$ as the relevance of document d_i with respect to topic t. REL is the set of the relevance degrees and it is a finite or countable totally ordered set which admits a minimum value rel_0. Thus, $r_t[d_i] = rel_0$ means that the document d_i is *not relevant* for the topic t. Higher values of REL stand for documents more and more relevant with respect to the topic t. To simplify the notation we denote $r_t[d_i]$ simply by $r[i]$.

The document utility model is defined by an infinite matrix of non negative real values $g(k, y)$, where $k \in \mathbb{N}$ and $y \in REL$. $g(k, y)$, called the **document utility loss**, defines the utility gained by a user when considering for the k–th time a document whose relevance is equal to y. We assume that the function $k \to g(k, y)$ is a **non increasing** function for any y. As a function of y we assume that $g(k, rel_0) = 0$ for any k and $y \to g(1, y)$ is a **non decreasing** function for any k. When $k > 1$ it is intuitive to stil assume that the function $y \to g(k, y)$ is **non decreasing**. However, it may be possible that a sensible utility model violates this assumption; for example the loss for visiting more times a highly relevant document might make it less relevant than the loss incurred by a partially relevant document visited the same number of times.

Example 4.1. *Continuing Example (3.1) and taking* $REL = \{rel_0, rel_1, rel_2, rel_3\}$ *a possible utility deriving model could be:*

$$g(k, rel_i) = i(1 - \lambda)^{k-1} \quad \textit{for all} \;\; 0 \leq i \leq 3 \;\; \textit{and} \; k \geq 1$$

where $\lambda \in [0, 1]$ *represents the percentage of utility lost at any visit. Taking* $\lambda = 0.5$, *the judged run* $[3, 2, 3, 0, 1]$ *and assuming that* $X_1 = 1$, $X_2 = 2$, $X_3 = 1$, $X_4 = 2$, $X_5 = 3$ *and* $X_6 = End$, *we get* $H = 5$ *and* $g(1, X_1) = 3$, $g(1, X_2) = 2$, $g(2, X_3) = 3 \times 0.5 = 1.5$, $g(2, X_4) = 2 \times 0.5 = 1$ *and* $g(1, X_5) = 3$.

5 THE P@H MEASURE

A run r is a (finite or countable) vector $(r[1], \ldots, r[N])$, where $r[i]$ has been defined above. Given a run r, we call **P@H** the measure

$$P@H(r) = \frac{1}{f(H)} \sum_{n=1}^{H} g\left(k(n), r[X_n]\right) \tag{2}$$

where $k(n) = |\{i \leq n : X_i = X_n\}|$ is the random number of times the user visited the rank position at which is after n steps and $h \to f(h)$ is a real positive non decreasing function.

Within *P@H*, we have:

- *browsing model*: it is defined by the transition matrix of the Markov chain $X = \{X_n, n \in \mathbb{N}\}$. Note that the transition probabilities can depend on the rank position, allowing us to model some notion of discount depending on the rank position, and/or on the relevance of the document at that rank position. In this way, we can frame both static and dynamic measures, as intended by Carterette [5].
- *document utility model*: it is determined by the document utility loss function, since $g(k(n), r[X_n])$ denotes the utility gained by the user when she visits the document at rank X_n for the $k(n)$–th time. In this way, we can account for the impact of how many times a document has been visited which, to the best of our knowledge, is new to our measure.
- *utility accumulation model*: it consists of the total utility collected by a user who moves according to one of our browsing models divided by a function of the (random) number of documents visited during her search, which act as a proxy of her effort. We have kept the utility accumulation model as simple as possible, even if still flexible enough to encompass many of the state-of-the-art evaluation measures. Note that, differently from "classical" evaluation measures which typically discount the accumulated utility by the rank position of the visited documents, we instead discount it by (a function f of) the stopping time, i.e. the stochastic number of steps actually performed by the user which is a more accurate estimation of her effort.

5.1 Ordering Runs

Since P@H is a random variable, in order to compare two different runs r and s, we need to define an order among random objects.

The first and relatively naive way to order runs is to compute the expectations of the two random variables, i.e. P@H for the run r and s, and order them accordingly to the *order of their expectations*:

$$\begin{gathered} P@H(r) \preceq_1 P@H(s) \\ \Updownarrow \\ \mathbb{E}\left[\frac{1}{f(H)}\sum_{n=1}^{H} g(k(n), r[X_n])\right] \leq \mathbb{E}\left[\frac{1}{f(H)}\sum_{n=1}^{H} g(k(n), s[X_n])\right] \end{gathered} \tag{3}$$

Note that the $\preceq_1$ order corresponds exactly to what is usually done when you order runs by a "classical" evaluation measure like AP. The pros are that this is a total order among runs while the cons are that we may underestimate the presence of outliers and/or other asymmetries that can greatly skew the expectations. Moreover, this order may be difficult to compute in a closed form due to the expectation of the division of two random variables, i.e. $\sum_{n=1}^{H}\ldots$ and $f(H)$; please, refer to the Electronic Appendix A for more details on this.

To make the explicit computation simpler, we can slightly modify the order $\preceq_1$ as follows:

$$\begin{gathered} P@H(r) \preceq_2 P@H(s) \\ \Updownarrow \\ \frac{\mathbb{E}\left[\sum_{n=1}^{H} g(k(n), r[X_n])\right]}{\mathbb{E}[f(H)]} \leq \frac{\mathbb{E}\left[\sum_{n=1}^{H} g(k(n), s[X_n])\right]}{\mathbb{E}[f(H)]} \end{gathered} \tag{4}$$

since the expectation of the two random variables $\sum_{n=1}^{H}\ldots$ and $f(H)$ is simpler to compute in closed form; please, refer to the Electronic Appendix B[1] for more details on this.

Despite these two orders are simple to understand, they are not really stochastic orders, where the whole distribution of P@H is taken into account. Thus, we can define the following stochastic order based on the notion of *stochastic dominance* [16, 31]:

$$\begin{gathered} P@H(r) \preceq_3 P@H(s) \\ \Updownarrow \\ \mathbb{P}\left[\frac{1}{H}\sum_{n=1}^{H} g(k(n), r[X_n]) > x\right] \leq \mathbb{P}\left[\frac{1}{H}\sum_{n=1}^{H} g(k(n), s[X_n]) > x\right] \\ \forall\, x \in \mathbb{R} \end{gathered} \tag{5}$$

The pros are that this order really embodies the idea that $P@H(s)$ is stochastically larger then $P@H(r)$, while the cons are that this is just a partial order and we have to decide how order runs in the case they cannot be compared, i.e. when there are swaps between the two probabilities for some values of x.

5.2 Simulation

We now compute our stochastic evaluation measures according to different browsing models by simulating the interaction of 100,000 users with respect to the runs $r = [1\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1]$ and $s = [0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0]$ of length $N = 10$. To ease the reproducibility of the experiments, the code for running the simulation is available at https://bitbucket.org/frrncl/ictir2020/.

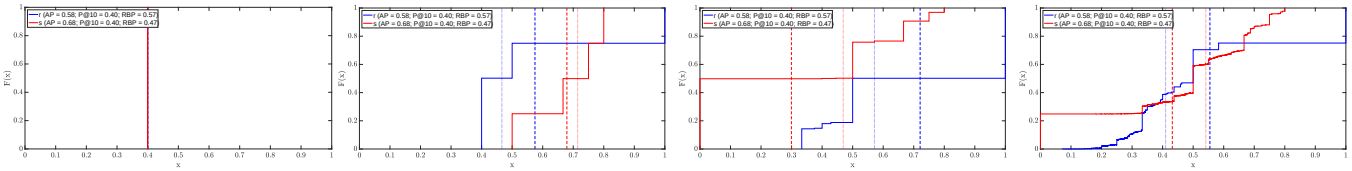
Figure 1 shows the results of the simulation and the cumulative distribution function $F(x)$ of the P@H measure for run r in blue and s in red. The dashed vertical line shows the estimated $\mathbb{E}\left[\frac{1}{f(H)}\sum_{n=1}^{H}\ldots\right]$, i.e. order $\preceq_1$; the dotted vertical line shows $\frac{\mathbb{E}\left[\sum_{n=1}^{H}\ldots\right]}{\mathbb{E}[f(H)]}$, i.e. order $\preceq_2$; note that the cumulative distribution function $F(x)$ of P@H allows us to visually estimate order $\preceq_3$, where the higher the value of $F(x)$, the smaller the run.

Figure 1a adopts a DFBM model which, as explained in Section 6.1, corresponds to Precision. Since we use a deterministic browsing model, the cumulative distribution function $F(x)$ collapses into a single value which also corresponds to the expectations computed according to orders $\preceq_1$ and $\preceq_2$. We can observe as these values correspond to the P@10 = 0.4 score reported in the legend of the plot. Note that, since both runs r and s have the same P@10 score the blue and red lines are overlapping; similarly, since orders $\preceq_1$ and $\preceq_2$ collapse into the same value, dashed and dotted lines overlap as well.

Figure 1b adopts a SFBM model which, as explained in Section 6.2, corresponds to AP. In particular, AP is equivalent to the expectation computed according to order $\preceq_1$ and, indeed, we can see that the dashed blue line corresponds to the AP = 0.58 score of the r run while the dashed red line corresponds to the AP = 0.68 score of the s run. Moreover, the plots of the cumulative distribution function $F(x)$ of P@H show richer information. Firstly, r and s are not comparable according to order $\preceq_3$, i.e. the stochastic dominance, since their distributions cross around $P@H = 0.75$. Secondly, even if the plain AP scores would lead to conclude that s is much better than r, the cumulative distribution function of P@H shows that, for roughly 20% of the users for who the two runs are performing quite well, i.e. P@H over 0.75, r is actually better than s. Finally, we can note how order $\preceq_2$ is consistent with order $\preceq_1$ still considering s better than r but in a more marked way.

Figure 1c adopts a SFBM model which, as explained in Section 6.3, corresponds to RBP with persistence parameter $p = 0.5$. In particular, RBP is equivalent to the expectation computed according to order $\preceq_2$ and, indeed, we can see that the dotted blue line corresponds to the RBP = 0.57 score of the r run while the dotted red line corresponds to the RBP = 0.47 score of the s run. In this case, we can observe as all the three orders agree since r is stochastic dominant over s according to order $\preceq_3$ and it is better than s also according to order $\preceq_2$, even if in a less marked way than according to order $\preceq_1$. Finally, as in the previous case, we can note how the cumulative distribution function of P@H provides richer information than the plain RBP score. For example, we can see that roughly half of the users has a uniform probability to observe low to medium performance for run s and medium to high performance for run s. This observation is consistent with the stop probability $s_i = 0.5$; indeed, the run s has a single cluster of relevant documents at the top of the ranking and users who will go deep down in the ranking will find no relevant documents, still spending time and effort in the run; conversely, r has the same number of relevant documents as s but scattered all over the ranking and thus persistent users are rewarded.

Figure 1d adopts a RWBM model which does not correspond to any "classical" evaluation measure. This RWBM model adopts the same forward probability $p_i = 0.50$ as RBP before but, differently



(a) DFBM model à la Precision; see Section 6.1 for further details. The transition probabilities are as follows: $p_1 = 1.00$; $p_i = 1.00$; $s_N = 1.00$.

(b) SFBM model à la AP; see Section 6.2 for further details. The transition probabilities depend on the relevance of the visited documents and on the total number of retrieved documents.

(c) SFBM model à la RBP with persistence $p = 0.5$; see Section 6.3 for further details. The transition probabilities are as follows: $p_1 = 0.50$ and $s_1 = 0.50$; $p_i = 0.50$ and $s_i = 0.50$; $s_N = 1.00$.

(d) RWBM model with a 25% document utility loss. The transition probabilities are as follows: $p_1 = 0.75$ and $s_1 = 0.25$; $p_i = 0.50$, $q_i = 0.25$, and $s_i = 0.25$; $q_N = 0.25$ and $s_N = 0.75$.

Figure 1: Comparison of two runs $r = [1\,0\,0\,1\,0\,0\,1\,0\,0\,1]$ and $s = [0\,1\,1\,1\,1\,0\,0\,0\,0\,0]$ using 100,000 users and different browsing models Each subplot shows the cumulative distribution function $F(x)$ of the P@H measure for run r in blue and s in red. The dashed vertical line shows order $\preceq_1$; the dotted vertical line shows order $\preceq_2$; note that the cumulative distribution function $F(x)$ of P@H allows us to visually estimate order $\preceq_3$, where the higher the value of $F(x)$, the smaller the run.

from the previous case, the stop probability is evenly split into a backward probability $q_i = 0.25$ and a stop probability $s_i = 0.25$; it also adopts a document utility loss equal to 25% at each subsequent visit of the same document. We can observe how this case behaves in a more complex way. Firstly, the two runs are not comparable according to order $\preceq_3$ but in a more mixed way, since their cumulative distribution functions $F(x)$ cross each other several times, suggesting that roughly 25%–30% of the users prefers s over r just in the medium performance range 0.40–0.68. Secondly, orders $\preceq_1$ and $\preceq_2$ do not agree and swap the two runs in quite a marked way.

6 CLASSICAL EVALUATION MEASURES

Since all the "classical" evaluation measures rely on a forward browsing model, either DFBM or SFBM, the user never visits the same document more than once. Therefore, we can assume that there is no utility loss, i.e. that $g(k, y) = g(y)$ for all k. Furthermore, since for any $n \leq N$ we have $X_n = n$, we write $r[n]$ instead of $r[X_n]$. Finally, in this section we assume that $REL = \{0, 1\}$ in the binary case and $REL = \{0, 1, \ldots, m\}$ in the multi-graded case.

6.1 Precision

To define precision, we adopt the DFBM model, since the user sequentially scans, in a deterministic way, all the rank positions up to the end of the run when she stops; we also take $g(y) = y$ and $f(h) = h$. Note that in this case $H = N$ almost surely. Therefore, we have:

$$P = \frac{1}{f(H)} \sum_{n=1}^{H} r[X_n] = \frac{1}{N} \sum_{n=1}^{N} r[n] \tag{6}$$

Note that, in this case, we have that $P = \mathbb{E}\left[\frac{1}{f(H)} \sum_{n=1}^{H} r[X_n]\right] = \frac{\mathbb{E}\left[\sum_{n=1}^{N} r[n]\right]}{\mathbb{E}[f(H)]}$, i.e. order $\preceq_1$ and order $\preceq_2$ are the same. Moreover, since in this case the distribution of P is a constant, they also coincide with order $\preceq_3$.

6.2 Average Precision

Let us assume binary relevance, i.e. $REL = \{0, 1\}$ and that $g(y) = y$, $f(h) = h$ as before. We also assume that (1) the user follows the SFBM model, since she sequentially scans the rank positions in a stochastic way, i.e. she may continue scanning or stop searching; (2) she moves forward with probability 1 after considering a not relevant document; (3) she may stop her search with a constant strictly positive probability only after considering a relevant document. This probability is equal to the reciprocal of the recall base RB_t, i.e. the total number of relevant documents for topic t.

Note that assumption (3) implies that the user moves forward with probability $1 - \frac{1}{RB_t}$ after considering a relevant document.

We can thus obtain AP as the expectation of the P@H measure using the above browsing model, i.e. AP coincides with order $\preceq_1$:

$$AP = \mathbb{E}\left[\frac{1}{H} \sum_{n=1}^{H} r[X_n]\right] = \frac{1}{RB_t} \sum_{i=1}^{N} r[i] \frac{1}{i} \sum_{j=1}^{i} r[j] \tag{7}$$

This interpretation of AP is correct only when the run retrieves all the RB_t relevant documents. Indeed, assumption (2) means that the probability of stopping at rank position i is $\mathbb{P}[H = i] = 0$ when the document at rank i is not relevant; assumption (3) means that the probability of stopping at rank position i is $\mathbb{P}[H = i] = \frac{1}{RB_t}$ when the document at rank i is relevant; in short we can write that $\mathbb{P}[H = i] = \frac{r[i]}{RB_t}$. Therefore, the total probability of stopping the search at any rank position is $\sum_{i=1}^{N} \mathbb{P}[H = i] = \sum_{i=1}^{N} \frac{r[i]}{RB_t}$ which sums to 1 only if the run retrieves all the RB_t relevant documents.

Therefore, the choice of $r[i]/RB_t$ is just an approximation of the exact value that is $r[i]/R_N$, where $R_N = r[1] + \ldots + r[N]$ is the total number of relevant documents actually retrieved by the run. Thus, the correct SFBM model uses the following transition probabilities

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \frac{r[1]}{R_N} & 0 & \frac{R_N - R_1}{R_N} & 0 & \cdots & 0 \\ \frac{r[2]}{R_N - R_1} & 0 & 0 & \frac{R_N - R_2}{R_N - R_1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{r[N-1]}{R_N - R_{N-2}} & 0 & 0 & 0 & \cdots & \frac{R_N - R_{N-1}}{R_N - R_{N-2}} \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

where $R_i = r[1] + \ldots + r[i]$.

Despite being deemed the de-facto gold standard measure in IR [32], AP is often criticised because of the lack of a convincing user model. Moffat and Zobel [20] argued that the model behind AP is abstract, complex, and far away from the real behavior of users interacting with an IR system, especially when it comes to its dependence on the recall base which is something actually unknown to real users. As a consequence, Robertson [22] proposed a simple but moderately plausibile user model: "*AP is the expected precision, given that our prediction of the user's stopping point is uniformly distributed over all the relevant documents for this topic*".

Our browsing model for AP further pushes Robertson's interpretation since assumption (3) sets the probability of stopping as the reciprocal of the recall base. However, apart this being an approximation of the correct stopping probability as discussed above, our browsing model further stresses how much the user model behind AP is not very realistic anyway. Indeed, even in the correct case without pretending to know the whole recall base, the probability of stopping the search depends on the total number of relevant documents retrieved by a run and this information is not known to the user before actually ending to scan the run.

6.3 Rank-Biased Precision (RBP)

RBP [20] assumes a user model where the user starts from the top ranked document and with probability p, called persistence, she goes to the next document or with probability $1-p$ stops. Therefore, the probability that the user stops her search at rank i is equal to $p^{i-1}(1-p)$. The RBP user model can be straightforwardly mapped to our SFBM model.

Assuming that the ranked result list is infinite, the expected number of document examined by the user is equal to $(1-p)^{-1}$, since

$$\sum_{i=1}^{\infty} i \cdot \left[p^{i-1} \cdot (1-p)\right] = \frac{1}{1-p}\,, \tag{8}$$

while the *total known expected utility derived by the user* is equal to $\sum_{i=1}^{N} r[i]p^{i-1}$; please, refer to the Electronic Appendix C for more details on this.

RBP is then defined as:

$$RBP = (1-p)\sum_{i=1}^{N} r[i]p^{i-1} \tag{9}$$

i.e. "*the total known expected utility derived by the user... divided by the average number of items inspected*", as stated by Moffat and Zobel.

In our framework, the intention of Moffat and Zobel in defining RBP is expressed as $\frac{\mathbb{E}\left[\sum_{n=1}^{H} r[X_n]\right]}{\mathbb{E}[f(H)]}$ with $g(y) = y$, $f(h) = h$, which corresponds to order $\preceq_2$. However, when the total number of retrieved documents N is **finite**, the above expression and eq. (8) are no more equal. Indeed, we cannot simply replace ∞ by N in the left hand side of (8), since the probability of stopping at rank N is in this case equal to p^{N-1}. Nevertheless, the computations in the Electronic Appendix C[2] show how the average number of documents visited in the case $N < \infty$ is equal to $\sum_{i=0}^{N-1} p^i = \frac{1-p^N}{1-p} =: K(p,N)$. As it can be expected, $K(p,N)$ converges to $(1-p)^{-1}$ as $N \to \infty$. In this sense, the value $(1-p)^{-1}$ derived in (8) is an approximation of the exact one $K(p,N)$ and to simplify the notation we define $K(p,\infty) = (1-p)^{-1}$. Therefore, following the intentions of the authors, the definition of RBP should be:

$$RBP(N) = \frac{\mathbb{E}\left[\sum_{n=1}^{H} r[X_n]\right]}{\mathbb{E}[H]} = K(p,N)^{-1}\sum_{n=1}^{N} r[i]p^{i-1}\,,$$

and the original *RBP* measure now becomes $RBP(\infty)$.

Note that Carterette [5] suggests an alternative interpretation of RBP as "*the expected relevance of the document at the stopping rank*". Even though this interpretation is correct, our approach is able to generalize RBP following the original intentions of the authors and to adjust it for not infinite rankings.

6.4 Discounted Cumulated Gain (DCG)

DCG [17] is a multi-graded relevance measure given by

$$\mathrm{DCG}_b(r) = \sum_{i=1}^{N} \frac{w(r[i])}{\max\{1, \log_b i\}} \tag{10}$$

where base b of the logarithm indicates the patience of the user in scanning the ranked result list and plays a role somewhat similar to the persistence parameter p of RBP. The weight function w monotonically maps *REL* into $[0, +\infty)$ to assign the weights corresponding to each relevance degree.

As explained before, what is classically considered as a discount factor in the utility accumulation model, in our framework becomes part of the browsing model and the distribution of its stopping time. Therefore, DCG can be seen as the expectation of a suitable P@H measure where the term $f(H)$ is a constant value, chosen equal to 1 for simplicity. Note that, in this case order $\preceq_1$ and $\preceq_2$ are the same. Also note that the constant $f(H)$ underlines how DCG is more focused on the utility accumulation process, i.e. the numerator of P@H, than on the effort required to derived such utility.

Electronic Appendix D[2] reports the detailed computations which lead to obtain DCG thanks to a FSBM model.

6.5 Expected Reciprocal Rank (ERR)

ERR [7] is a multi-graded relevance measure given by

$$\mathrm{ERR}(r) = \sum_{i=1}^{N} \frac{1}{i}\prod_{j=1}^{i-1}(1-R_j)R_i \tag{11}$$

where R_i denotes the probability that the user stops her search after considering the document at rank position i, since she is satisfied. Usually R_i depends on the relevance of the document at rank position i, but does not depend on the rank position itself. Chapelle et al. [7] assume $R_i = \rho(r[i])$ where $\rho(j) = \frac{2^j-1}{2^k}$.

We can frame ERR into P@H by using a "quantised" utility accumulation model, since here the interest is in the "effort" of the user for achieving a positive utility rather than in the amount of such utility. To this end, g has to depend also on n and we assume that $g_1(y) = 1$ for any k, while $g_n \equiv 0$ for $n > 1$. Furthermore, since we do not have an additional state for the case of a user ending her search "unsatisfied", we have that, once the user arrives at rank N, she stops her search with probability 1, defining therefore $R_N = 1$. Taking $f(y) = y$, and defining the FSBM model with transition

matrix:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ R_1 & 0 & 1-R_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{N-1} & 0 & 0 & \cdots & 1-R_{N-1} \\ R_N & 0 & 0 & \cdots & 0 \end{bmatrix}$$

we obtain

$$\text{ERR(r)} = \mathbb{E}\left[\frac{1}{H}\right] = \sum_{i=1}^{N} \frac{1}{i}\,\mathbb{P}[H=i] = \sum_{i=1}^{N} \frac{1}{i} \prod_{j=1}^{i-1} (1-R_j)R_i$$

since $\mathbb{P}[H=i] = \prod_{j=1}^{i-1}(1-R_j)R_i$.

7 CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a new family of stochastic evaluation measures which rely on Markov chains and the distribution of their stopping time H to overcome several limitations of "classical" evaluation and to extend the general framework proposed by Carterette [5]. In particular, our P@H measure combines the simplicity and intuitiveness behind Precision, i.e. a straightforward utility accumulation model where the total derived utility is discounted by the effort incurred to gain it, with a powerful browsing model, where the way in which the user interacts with the ranked result list is captured by a Markov chain. We have shown, both theoretically and experimentally, how these new measures are actually random variables associated with a whole distribution of scores rather than a single deterministic value and that they provide a richer description of the performance range experienced by users. Moreover, we have demonstrated, both theoretically and experimentally, how state-of-the-art evaluation measures – namely Precision, AP, RBP, DCG, and ERR – can all be coherently represented in our new framework. Finally, we have shown that our new family of measures can be used to simulate how users interact with system runs and how to score their performance accordingly.

We have described how these new measures can be leveraged to define different ways of comparing and ordering system runs and we have experimentally shown that these orders may or may not agree with each other and that they may be just partial orders, being some run pairs not comparable. Future work will thus concern a deeper investigation of these orders, under which conditions they are possible, and when they agree or not.

Moreover, Ferrante et al. [14] have recently proposed a theory for the interval scale properties of deterministic evaluation measures which is based on how measures order runs. Future work will thus concern a deeper investigation of these orders, under which conditions they are possible, and when they agree or not. We will investigate how our orders can be related to the orders introduced by Ferrante et al. [14] in their recent theory about evaluation measures and how we can constraint P@H to ensure that the generated measures are interval scales within the framework of the probabilistic approach to the measurement scales [23].

Finally, we will investigate additional Markovian browsing models, able to grasp more complex user interactions, like reading a snippet or clicking a document, in order go also towards the user-oriented side of the evaluation spectrum.

REFERENCES

[1] L. Azzopardi, P. Thomas, and N. Craswell. 2018. Measuring the Utility of Search Engine Result Pages. *SIGIR 2018*, 605–614.
[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2015. User Variability and IR System Evaluation. *SIGIR 2015*, 625–634.
[3] F. Baskaya, H. Keskustalo, and K. Järvelin. 2013. Modeling Behavioral Factors in Interactive Information Retrieval. *CIKM 2013*, 2297–2302.
[4] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, USA, 53–78.
[5] B. A. Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. *SIGIR 2011*, 903–912.
[6] B. A. Carterette, E. Kanoulas, and E. Yilmaz. 2012. Incorporating Variability in User Behavior into Systems Based Evaluation. In *CIKM 2012*, 135–144.
[7] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *CIKM 2009*, 621–630.
[8] W. S. Cooper. 1968. Expected Search Length: A Single Measure of Retrieval Effectiveness. *American Documentation* 19, 1, 30–41.
[9] S. Dungs and N. Fuhr. 2017. Advanced Hidden Markov Models for Recognizing Search Phases. In *ICTIR 2017*, 257–260.
[10] M. Ferrante, N. Ferro, and E. Losiouk. 2020. How do interval scales help us with better understanding IR evaluation measures? *IRJ* 23, 3, 289–317.
[11] M. Ferrante, N. Ferro, and M. Maistro. 2014. Injecting User Models and Time into Precision via Markov Chains. *SIGIR 2014*, 597–606.
[12] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. *ICTIR 2015*, 21–30.
[13] M. Ferrante, N. Ferro, and S. Pontarollo. 2017. Are IR Evaluation Measures on an Interval Scale?. *ICTIR 2017*, 67–74.
[14] M. Ferrante, N. Ferro, and S. Pontarollo. 2019. A General Theory of IR Evaluation Measures. *TKDE* 31, 3, 409–422.
[15] N. Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3, 32–41.
[16] J. Hadar and W. R. Russell. 1969. Rules for Ordering Uncertain Prospects. *The American Economic Review* 59, 1, 25–34.
[17] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *TOIS* 20, 4, 422–446.
[18] D. Maxwell and L. Azzopardi. 2016. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *SIGIR 2016*, 1141–1144.
[19] D. M. Maxwell. 2019. *Modelling Search and Stopping in Interactive Information Retrieval*. Ph.D. Dissertation. University of Glasgow, Scotland, UK.
[20] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *TOIS* 27, 1, 2:1–2:27.
[21] J. R. Norris. 1998. *Markov chains*. Cambridge University Press, UK.
[22] S. Robertson. 2008. A New Interpretation of Average Precision. *SIGIR 2008*, 689–690.
[23] G. B. Rossi. 2014. *Measurement and Probability. A Probabilistic Theory of Measurement with Applications*. Springer-Verlag, USA.
[24] T. Sakai and Z. Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. *SIGIR 2013*, 473–482.
[25] M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *FnTIR* 4, 4, 247–375.
[26] P. Serdyukov, N. Craswell, and G. Dupret. 2012. WSCD2012: Workshop on Web Search Click Data 2012. *WSDM 2012*, 771–772.
[27] M. D. Smucker and C. L. A. Clarke. 2012. Stochastic Simulation of Time-Biased Gain. In *CIKM 2012*, 2040–2044.
[28] M. D. Smucker and C. L. A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. *SIGIR 2012*, 95–104.
[29] P. Thomas, P. Bailey, A. Moffat, and F. Scholer. 2014. Modeling Decision Points in User Search Behavior. *IIiX 2014*, 239–242.
[30] D. van Dijk, M. Ferrante, N. Ferro, and E. Kanoulas. 2019. A Markovian Approach to Evaluate Session-based IR Systems. I*ECIR 2019*, 621–635.
[31] J. von Neumann and O. Morgenstern. 1953. *Theory of Games and Economic Behavior* (3rd ed.). Princeton University Press, USA.
[32] E. Yilmaz and J. A. Aslam. 2006. Estimating Average Precision With Incomplete and Imperfect Judgments. *CIKM 2006*, 102–111.
[33] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. 2010. Expected Browsing Utility for Web Search Evaluation. *CIKM 2010*, 1561–1565.
[34] F. Zhang, Y. Liu, X. Li, M. Zhang, Y. Xu, and S. Ma. 2017. EvaluatingWeb Search with a Bejeweled Player Model. *SIGIR 2017*, 425–434.
[35] Y. Zhang, X. Liu, and C. Zhai. 2017. Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation. *ICTIR 2017*, 193–200.

Exploiting Stopping Time to Evaluate Accumulated Relevance

Electronic Appendix

Marco Ferrante
Department of Mathematics "Tullio Levi-Civita",
University of Padua, Italy
ferrante@math.unipd.it

Nicola Ferro
Department of Information Engineering,
University of Padua, Italy
ferro@dei.unipd.it

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**.

KEYWORDS

evaluation measure; Markov chain; stopping time; user model

ACM Reference Format:
Marco Ferrante and Nicola Ferro. 2020. Exploiting Stopping Time to Evaluate Accumulated Relevance: Electronic Appendix. In *The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3409256.3409832

In this appendices we present some explicit computations regarding the total orders defined in the paper and how we derive the expressions for some classical measures. For simplicity, we consider $REL = \{0, 1\}$ in the binary case and $REL = \{0, 1, \dots, m\}$ in the multi-graded case.

A COMPUTATION OF ORDER $\preceq_1$

To be able to organize runs according to the total order $\preceq_1$, we need to compute $\mathbb{E}[P@H] = \mathbb{E}[\frac{1}{f(H)} \sum_{n=1}^{H} g(k(n), r[X_n])]$.

In the case of the SFBM model with constant transition probabilities $p_i \equiv p$ and taking $f(y) = y$ and $g(k, y) = y$, we obtain:

$$\mathbb{E}[P@H] = \sum_{h=1}^{N-1} p^{h-1}(1-p)\frac{1}{h}\sum_{i=1}^{h} g(1, r[i]) + p^N \frac{1}{N}\sum_{i=1}^{N} g(1, r[i])$$

which is similar to RBAP (Rank based Average Precision) defined by Carterette [1].

However, for a general transition matrix $P = \{p_{ij}\}_{i,j\in\mathcal{S}}$, as we have in the RWBM model, we obtain the following expression

$$\begin{aligned}\mathbb{E}_i[P@H] &= \mathbb{E}_i\left[\frac{1}{H}\sum_{n=1}^{H} g(k(n), r[X_n])\right] \\ &= \sum_{h=1}^{\infty}\sum_{i_j\neq End, j\leq h}\left(\frac{1}{h}\sum_{n=1}^{h} g(k(n), r[i_n])\right) p_{i,i_2}\cdots p_{i_h,End}\end{aligned}$$

which cannot be further simplified.

ICTIR '20, September 14–17, 2020, Virtual Event, Norway

ACM ISBN 978-1-4503-8067-6/20/09…$15.00
https://doi.org/10.1145/3409256.3409832

B COMPUTATION OF ORDER $\preceq_2$

To be able to order runs according to the total order $\preceq_2$„ we need to compute two expectations, namely $\mathbb{E}[\sum_{n=1}^{H} g(k(n), r[X_n])]$ and $\mathbb{E}[H]$. Since $H = \sum_{n=1}^{H} g(k(n), r[X_n])$ if $g(k, y) \equiv 1$, it is sufficient to evaluate the first one in order to also derive the second. Let $T(H) = \sum_{n=1}^{H} g(k(n), r[X_n])$ and assume that $g(k, y) = g(y)$ for any k (relevant just in the case of the Random Walk model). For any $i \in \mathcal{N}$

$$\begin{aligned}&\mathbb{E}\left[T(H)|X_1 = i\right] = \mathbb{E}_i[f(H)] = \mathbb{E}_i\left[\textstyle\sum_{n=1}^{H} g(r[X_n])\right] \\ &= \textstyle\sum_{j\in\mathcal{S}} \mathbb{E}_i\left[\sum_{n=1}^{H} g(r[X_n])\middle| X_2 = j\right] p_{ij} \\ &= g(r[i]) + \textstyle\sum_{j\in\mathcal{N}} \mathbb{E}_j\left[T(H)\right].\end{aligned}$$

If $E = [\mathbb{E}_1[T(H)], \dots, \mathbb{E}_N[T(H)]]^{tr}$, where ^{tr} means taking the transpose, and $\rho = (r[1], \dots, r[N])^{tr}$ we have in matrix notation

$$E = (Id - \widetilde{P})^{-1}\rho, \tag{1}$$

where Id is the identity matrix and $\widetilde{P} = \{p_{ij}\}_{i,j\in\mathcal{N}}$.

By a similar computation, we can derive the conditional variances $\mathrm{Var}_i[T(H)] = \mathbb{E}_i[T^2(H)] - (\mathbb{E}_i[T(H)])^2$. Defined

$$V = [\mathrm{Var}_1[T(H)], \dots, \mathrm{Var}_N[T(H)]]^{tr}$$

and the diagonal matrix $R = \mathrm{diag}\,(r[1], \dots, r[N])$, we have that

$$V = (Id - \widetilde{P})^{-1} R\left[Id - 2\widetilde{P}(1 - \widetilde{P})^{-1}\right]\rho,$$

To conclude, in this case we are also able to evaluate the conditional distribution of $T(H)$ in a closed form. Let $s_i(n) \coloneqq \mathbb{P}_i(T(H) = n)$ and denote $s = (s_1(n) \dots, s_i(n))^{tr}$; a computation similar to the previous ones gives

$$\begin{aligned}s = &\left[\left[Id - (Id - R)\widetilde{P}\right]^{-1} R\widetilde{P}\right]^{n-1} \\ &\times\left[Id - \left[Id - (\mathbf{1} - R)\widetilde{P}\right]^{-1} R\widetilde{P}\right]\left[Id - (Id - R)\widetilde{P}\right]^{-1}\rho,\end{aligned}$$

From these computations and taking $g(y) \equiv 1$ we derive the expectation, variance and distribution of H alone. Indeed, we get

$$E = (Id - \widetilde{P})^{-1} e, \tag{2}$$

where $e = (1, \dots, 1)^{tr}$,

$$V = (Id - \widetilde{P})^{-1}\left[Id - 2\widetilde{P}(1 - \widetilde{P})^{-1}\right] e$$

and

$$s = \widetilde{P}^{n-1}\left[Id - \widetilde{P}\right] e\ .$$

In the particular case of the Stochastic forward model with $p_i \equiv p$ and g not depending on $k(n)$, as natural in this case, we can compute explicitly $\mathbb{E}_1[T(H)]$ and $\mathrm{Var}_1[T(H)]$ (which are the most

interesting quantities, since we always assume that the search starts from the first document with probability 1). Indeed we have

$$\mathbb{E}_1[T(H)] = \sum_{i=1}^{N} r[i]p^{i-1},$$

and

$$\mathrm{Var}_1[T(H)] = \sum_{m=2}^{N} r[m]p^{m-1}(1-p^{m-1}) \times \left[r[m] + 2 \sum_{i=m+1}^{N} r[i]p^{i-m} \right].$$

Note that $\mathbb{E}_1[T(H)]$ coincides, up to a constant, with RBP.

C RBP AS $\mathbb{E}$[P@H]

Consider the RWBM model, in the simplified case of $p_i = p$ and $q_i = q$ for any i, with $0 < p, q, 1-p-q < 1$ to avoid corner cases. We can interpret this as a generalization of *Rank-Biased Precision (RBP)* and we are interested to perform the computation of the two expectations in the total order $\preceq_2$.

Assuming as before that $g(k, y) = y$ and that $g(0) = 0$, while $g(i) = 1$ for any $i \geq 1$ (that includes the classic binary relevance models), the explicit computation of $\mathbb{E}_1[\sum_{n=1}^{H} r[X_n]]$ is difficult to perform and depends on the specific run considered. For example, taking $N = 6$ and the run $r = (1, 0, 0, 1, 0, 1)$, we get

$$\mathbb{E}_1[P@H] = \frac{1 - 4pq + p^3 + 3p^2q^2 - p^4q + p^5}{1 - 5pq + 6p^2q^2 - p^3q^3}$$

On the contrary, it is possible to derive the following explicit expression

$$\mathbb{E}_1[H] = A\frac{\sqrt{1-4pq}}{p} + \frac{2p - 1 + \sqrt{1-4pq}}{2p(1-p-q)} \tag{3}$$

where

$$A = \left(\frac{1}{1-p-q} g_1(p, q, N) - \frac{p}{1-p-q} \right) \times \Big(g_2(p, q, N) - g_1(p, q, N) \Big)^{-1},$$

and

$$g_1(p, q, N) = \left(\frac{1 - \sqrt{1-4pq}}{2p} \right)^N - q \left(\frac{1 - \sqrt{1-4pq}}{2p} \right)^{N-1}$$

$$g_2(p, q, N) = \left(\frac{1 + \sqrt{1-4pq}}{2p} \right)^N - q \left(\frac{1 + \sqrt{1-4pq}}{2p} \right)^{N-1}.$$

Denoting $\mathbb{E}_1[H] = K(p, q, N)$, we get

$$K(p, 0, N) = \frac{1 - p^N}{1 - p}.$$

If we are interested to use this number for a normalization purpose, the fact that it depends on a possible final "tail" of non relevance documents could be a potential problem. In the spirit of RBP, it is therefore possible to substitute this by its limit for N that goes to infinity, which we will denote by $K(p, q, \infty)$. A simple computation gives

$$K(p, q, \infty) = \frac{1}{1-p-q} \left(\frac{2p - 1 + \sqrt{1-4pq}}{2p} \right)$$

Note that $K(p, 0, \infty) = (1-p)^{-1}$.

D DCG AS $\mathbb{E}$[P@H]

Let us interpret

$$DCG_b(r) = \sum_{i=1}^{N} \frac{w(r[i])}{\max\{1, \log_b i\}}$$

as the expectation of a suitable P@H measure. Consider a FSBM model where the forward transition probability p_i is not constant, and evaluate

$$\mathbb{E}\Big[\sum_{m=1}^{H} w(r[X_m]) \Big] = \sum_{n=1}^{N} \Big(\sum_{i=1}^{n} w(r[i]) \Big) \mathbb{P}[H = n]$$

$$= \sum_{i=1}^{N} \Big(\sum_{n=i}^{N} \mathbb{P}[H = n] \Big) w(r[i])$$

Then we have to solve the linear system

$$\frac{1}{\max\{1, \log_b i\}} = \sum_{n=i}^{N} \mathbb{P}[H = n]$$

for $i \in \{1, \ldots, N\}$. Taking for instance $b = 2$, we obtain

$$1 = \sum_{n=1}^{N} \mathbb{P}[H = n]$$

which is trivially true; then

$$1 = \sum_{n=2}^{N} \mathbb{P}[H = n]$$

which implies that $\mathbb{P}[H = 1] = 0$,

$$\frac{1}{\log_2 3} = \sum_{n=3}^{N} \mathbb{P}[H = n]$$

which implies that $\mathbb{P}[H = 2] = \frac{\log_2 3 - 1}{\log_2 3}$,

$$\frac{1}{\log_2 i} = \sum_{n=i}^{N} \mathbb{P}[H = n],$$

which implies $\mathbb{P}[H = i] = \frac{\log_2(i+1) - \log_2(i)}{\log_2(i+1)\log_2(i)}$, for $i \in \{3, \ldots, N-1\}$, and finally

$$\mathbb{P}[H = N] = \frac{1}{\log_2 N}.$$

Therefore, leveraging the FSBM model, we obtain that

$$\mathbb{P}[H = n] = (1 - p_n) \prod_{i=1}^{n-1} p_i$$

for $n \in \{1, \ldots, N-1\}$, while

$$\mathbb{P}[H = N] = \prod_{i=1}^{N-1} p_i.$$

At the end we obtain the following transition probabilities $p_1 = 1$ and $p_i = \frac{\log_2(i)}{\log_2(i+1)}$ for $i \in \{2, \ldots, N-1\}$ which correspond to the

transition matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 1-\frac{\log_2 2}{\log_2 3} & 0 & 0 & \frac{\log_2 2}{\log_2 3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1-\frac{\log_2 (N-1)}{\log_2 N} & 0 & 0 & 0 & \cdots & 0 & \frac{\log_2 (N-1)}{\log_2 N} \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

REFERENCES

[1] B. A. Carterette. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft, editors, *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 903–912. ACM Press, New York, USA, 2011.