Injecting User Models and Time into Precision via Markov Chains

Marco Ferrante
Dept. Mathematics
University of Padua, Italy
ferrante@math.unipd.it

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

Maria Maistro
Dept. Information Engineering
University of Padua, Italy
maistro@dei.unipd.it

ABSTRACT

We propose a family of new evaluation measures, called *Markov Precision (MP)*, which exploits continuous-time and discrete-time Markov chains in order to inject user models into precision. Continuous-time MP behaves like time-calibrated measures, bringing the time spent by the user into the evaluation of a system; discrete-time MP behaves like traditional evaluation measures. Being part of the same Markovian framework, the time-based and rank-based versions of MP produce values that are directly comparable.

We show that it is possible to re-create average precision using specific user models and this helps in providing an explanation of *Average Precision (AP)* in terms of user models more realistic than the ones currently used to justify it. We also propose several alternative models that take into account different possible behaviors in scanning a ranked result list.

Finally, we conduct a thorough experimental evaluation of MP on standard TREC collections in order to show that MP is as reliable as other measures and we provide an example of calibration of its time parameters based on click logs from Yandex.

Categories and Subject Descriptors

H.3.4 [**Information Search and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Evaluation; Markov Precision; User Model; Time

1. INTRODUCTION

Experimental evaluation has been central to *Information Retrieval (IR)* since its beginning [15] and Cranfield is the predominant paradigm for carrying out system-oriented experimentation [11]. Over the decades, several measures have been proposed to evaluate retrieval effectiveness.

AP [5] represents the "gold standard" measure in IR [35], known to be stable [3] and informative [1], with a natural top-heavy bias and an underlying theoretical basis as approximation of the area under the precision/recall curve. Nevertheless, due to its dependence on the recall base, it assumes a perfect knowledge of the relevance of each document in the collection, which is an approximation when pooling is adopted and not assessed documents are assumed to be not relevant [14], and is even more exacerbated in the case of large scale or dynamic collections [4, 35].

However, the strongest criticism to AP comes from the absence of a convincing user model for it, a feature which is deemed extremely important in order to make the interpretation of a measure meaningful and to bridge the gap between system-oriented and user-oriented studies [7, 21, 31]. In this respect, [22] argued that the model behind AP is abstract, complex, and far 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, [25] proposed a simple but moderately plausibile user model for AP, which allows for a mix of different behaviors in the population of users.

In this paper, we take up from the final considerations of [25], at page 690: "this argument could provide the basis for a more elaborate model, by for example basing the set of $p_s(n)$ on some more sophisticated view of stopping behaviour", where $p_s(n)$ is the probability that the user satisfaction point is the document at rank n.

We propose a family of measures of retrieval effectiveness, called *Markov Precision (MP)*, where we exploit Markov chains [23] to inject different user models into precision and which does not depend on the recall base. We represent each position in a ranked result list with a state in a Markov chain and the different topologies and transition probabilities among the states of the Markov chain allow us to model the different and perhaps complex user behaviors and paths in scanning a ranked result list. The invariant distribution of the Markov chain provides us with the probability of the user being in a given state/rank position in stationary conditions and we use these probabilities to compute a weighted average of precision at those rank positions.

The framework we propose is actually more general and it is based on continuous-time Markov chains in order to take into account also the time a user spends in visiting a sin-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609637.

gle document. It is then possible to extract a discrete-time Markov chain, when considering only the transitions among rank positions and not the time spent in each document. This gives us a two-fold opportunity: when we consider the discrete-time Markov chain, we are basically reasoning as traditional evaluation measures which assess the utility for the user in scanning the ranked result list; when we consider the continuous-time Markov chain,we also embed the information about the time spent by the user in visiting a document and we have a single measure including both aspects. This represents a valuable contribution of the paper since, up to now, rank and time have been two separate variables according to which retrieval effectiveness is evaluated [31].

The Markov chain approach relies on some assumptions – e.g. no long-term memory and exponentially distributed holding times – which may seem oversimplifications of the reality, e.g. a user who considers the whole history of visited documents to decide whether to stop or not. However, other measures, such as *Rank-Biased Precision (RBP)* [22] where transitioning to the next document or stopping is a step-by-step decision based just on the persistence parameter, are memory-less in this sense. Moreover, a Markovian model is simple enough to be easily dealt with while still being quite powerful and this work intends to be a first step towards a richer world of models that we will explore in the future.

We then propose some basic models for the transition matrix of the Markov chain. Clearly, this is not intended to be an exhaustive list of all the possible models but more of an exemplification of how it is possible to plug different user models into the framework. Still, these basic models provide a second valuable contribution of the paper. Indeed, we will show how some of these models, when provided with the same level of information about the recall base as AP, actually are AP, thus giving an explanation of it in terms of a slightly richer user model than the one of [25]. We will also show how some of them are extremely highly correlated to AP, thus suggesting how AP can be considered a very good approximation of more complex user strategies. This helps in shedding some light on why AP is the de-facto "gold standard" in IR, even though it has been so often criticized.

Finally, we conduct a thorough experimental evaluation of the MP measure both using standard *Text REtrieval Conference (TREC)*[1] collections and click-logs with assessed queries made available by Yandex [29]. The results show that MP is comparable to other measures for some desirable properties like robustness to pool downsampling while the Yandex click-logs allow us to estimate the time spent by the users on the documents and apply the continous-time Markov chain.

The paper is organized as follows: Section 2 presents the related works; Section 3 discusses other pertinent measures to which MP will be compared; Section 4 fully introduces MP; Section 5 reports the conducted experimental evaluation of MP; and Section 6 draws some conclusion and provides an outlook for future work.

2. RELATED WORKS

Markov-based approaches have been previously exploited in IR, for example: Markov chains have been used to generate query models [19], for query expansion [12, 20], and for document ranking [13]. However, to the best of our knowledge, Markov chains have not been applied to the definition of a fully-fledged measure for retrieval effectiveness.

[8] uses Markov chains to address the placement problem in the case of two-dimensional results presentation: they have to allocate images on a grid to maximize the expected total utility of the user, according to some evaluation measure, and the Markov chain models how the user moves in the grid. Their approach differs from ours since they are not defining a measure of effectiveness which embeds a Markov chain but they rather solve an optimization problem via a Markov chain; moreover, they only use discrete-time Markov chains and limit transitions only to adjacent states. What we share is the idea that a Markov chain can be used to model how a user scans a result list, mono dimensional in our case, two-dimensional in their case.

When it comes to other evaluation measures, the focus of the paper is on lab-style evaluation, search tasks with informational intents [2], and binary relevance. So, for example, measures for novelty and diversity are out of the scope of the present paper [10] as are measures for graded relevance like *Discounted Cumulated Gain (DCG)* [16], *Expected Reciprocal Rank (ERR)* [6], or Q-measure [26].

With regard to the time dimension brought in by the continuous-time Markov chain, the most relevant work is *Time-Biased Gain (TBG)* [30, 31]. We share the idea of getting time into evaluation measures but we adopted a different approach. While TBG substitutes traditional evaluation measures, MP provides a single framework for keeping both aspects depending on which Markov chain you use. With respect to the user model adopted in TBG, there are some relevant differences: first, we use full Markov models while [30] at page 2014 points out that "our model can be viewed as a semi-Markov model"; then, TBG assumes a sequential scanning of the result lists where MP allows the user to move and jump backward and forward in the results list. What TBG addressed and is not in the scope of the present work is how to calibrate the measure with respect to time: [31] proposed a procedure to calibrate time with respect to document length and [30] extended it to stochastic simulation. In the present work, we provide a basic example of calibration based on the estimation of average time spent per document from click logs, just to show how the parameters of the framework could be tuned. However, in the future, nothing prevents us (or others) from investigating more advanced calibration strategies or applying those proposed by [30, 31].

Previous work on click logs [17] has reported that, on average, users scan ranked list in a forward linear fashion while MP allow users to move forward and backward in a ranked list. As reported in Section 5.5, from Yandex logs, we found that 21% of the users move backward in the ranked list, thus supporting our assumption, even if more exploration on this is left for future work. Moreover, U-measure [28] is a recent proposal which shares with MP the idea of removing the constraint of the linear scan but it does not adopt Markov models and has also somewhat different goals, such as evaluating complex tasks like multi-query sessions and diversified IR.

When it comes to other ways of modelling user behaviour into evaluation measures, [7] proposes relying on three components: a browsing model, a model of document utility, and a utility accumulation model. Even if we took up from [25], MP can also be framed in the light of the work of [7]. Indeed,

[1] `http://trec.nist.gov/`

the Markovian model provides us with the browsing model, precision account for the model of document utility, and the weighted average of precision by the invariant distribution of the Markov chain supplies the utility accumulation model.

Thus, evaluation measures of direct comparison, which will be detailed in Section 3, are those built around the concept of precision, namely AP, P@10, and Rprec [5]. RBP [22] comes into play as a binary evaluation measure not dependent on the recall base, even though it is not built around the concept of precision despite its name. Finally, we are also interested in *Binary Preference (bpref)* [4], just to have a comparison point when testing MP with respect to reduced-size pools. In this last respect, we are not interested in infAP [35], since we are neither looking for an estimator of AP nor investigating alternative strategies for pool down-sampling. For the same reason, we are not interested here in experimenting with respect to condensed-list measures [27].

3. OTHER EVALUATION MEASURES

Let us consider a ranked list of T documents in response to a given topic, let d_n be the document retrieved at position $n \leq T$ whose relevance is denoted by a_n, equal to 1 if the document is considered relevant and 0 otherwise. The ranked list of documents is denoted with $\mathcal{D} = \{d_i, i \leq T\}$ and $\mathcal{R} = \{i_j : j = 1, \ldots, T \text{ and } a_{i_j} = 1\}$ is the set of the ranks of the relevant documents, whose cardinality is $r = |\mathcal{R}|$ and which indicate the total number of relevant retrieved documents by the system for the given topic. Let RB be the recall base of the topic, i.e. the total number of judged relevant documents for a given topic, and NRB the total number of judged not relevant documents for a given topic.

The *precision at rank* n is thus defined as

$$\text{Prec}(n) = \frac{1}{n} \sum_{m=1}^{n} a_m \tag{1}$$

which corresponds to the percentage or "density" of relevant documents present among the first n, n included, in the list. Note that Rprec is Prec(RB), which makes clear its dependence on the recall base.

The *recall at rank* T is defined as

$$\text{Rec}(T) = \frac{r}{RB} \tag{2}$$

which corresponds to the fraction of relevant documents of the specific run with respect to the total number of judged relevant documents.

3.1 Average Precision (AP)

The original definition of *Average Precision (AP)* [5] is the average over all RB judged relevant documents of the precision at their ranks, considering zero the precision at the not retrieved relevant documents:

$$AP = \frac{1}{RB} \sum_{i \in \mathcal{R}} \text{Prec}(i) = \frac{r}{RB} \cdot \frac{1}{r} \sum_{i \in \mathcal{R}} \text{Prec}(i) \tag{3}$$

where, in the last equation, the first operand is the recall and the second one is the arithmetic mean of the precisions at each relevant retrieved document. This formulation further highlights the dependence of AP on the recall base and the recall itself.

As previously discussed, [25] proposed a simple, probabilistic user model measure of effectiveness called *Normalized Cumulative Precision (NCP)*, which includes AP as a particular case. The author assumes that any given user will stop his search at a given document in the ranked list, that we call its satisfaction point, according to a common probability law.

Furthermore, he considers that a user will stop his search only at relevant documents and that the probability that he stops at any given relevant documents is fixed and independent from the specific run he is considering, while it is 0 at any non relevant document. So, he defines a probability distribution p_s on the set of all the documents available for a given topic.

Given a specific run and the set of its retrieved documents $\mathcal{D}$, the definition of the NCP is then the expectation (average) of the precision at the ranks of the retrieved, relevant documents, accordingly to a distribution $p_s(\cdot)$, i.e.

$$NCP(p_s) = \mathbf{E}_{p_s}[\text{Prec}(n)] = \sum_{n=1}^{+\infty} p_s(d_n)\text{Prec}(n) \ .$$

It is easy to see that the above definition of AP is in this context equal to the NCP measure when we choose the uniform law p_U over all the relevant documents for the topic

$$p_U(d_n) = \begin{cases} \dfrac{1}{RB} & \text{if } d_n \text{ is relevant} \\ 0 & \text{otherwise} \end{cases}$$

The previous user model is simple and it can be considered as a starting point for more sophisticated models, as also suggested by [25] itself. As in the case of AP, the assumption that the user knows the recall base of a given topic is a weakness of this model. Furthermore, the probability that a user stops their search at a given document on a specific run depends on a probability distribution defined on the whole set of relevant documents available for a given topic.

The choice of the uniform distribution to determine the stopping point in a given search is itself of difficult interpretation, since this means that any relevant document in a ranked list of retrieved documents has the same probability.

We will see in the next section how, stepping from the intuition behind NCP, we can define, thanks to simple Markov chains, a more realistic user model, how AP can be still considered as a good approximation in many cases and how to generalize AP to a whole new class of Markovian models.

3.2 Rank-Biased Precision (RBP)

Rank-Biased Precision (RBP) [22] assumes a user model where the user starts from the top ranked document and with probability p, called persistence, goes to the next document or with probability $1 - p$ stops. RBP is defined as follows:

$$RBP = (1 - p) \sum_{i \in \mathcal{R}} p^{i-1} \tag{4}$$

It can be noted that, despite its name, RBP does not depend on the notion of precision. Nevertheless, it represents a measure for binary relevance which does not depend on the recall base and thus gives a comparison point in this last respect for MP.

3.3 Binary Preference (bpref)

Binary Preference (bpref) [4, 32] is a measure based on binary preferences and it evaluates systems using only the judged documents. It can be thought of as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones:

$$bpref = \frac{1}{RB} \sum_{i \in \mathcal{R}} \left(1 - \frac{|j \text{ ranked higher than } i|}{\min(RB, NRB)} \right) \quad (5)$$

where j is a member of the first RB not relevant retrieved documents. bpref has proved to be quite robust in the case of incomplete and imperfect relevance judgements. Here, for us, it represents a comparison point when evaluating MP with respect to reduced-size pools.

It can be noted how heavily bpref depends on the recall base RB. This is not only a scale factor as in the case of AP but it also determines the cardinality of the set from which the not relevant documents j are taken. Moreover, it makes use also of NRB, the total number of judged not relevant documents, a kind of information which is hard to imagine available to any real user. So, in a sense, it seems much more a "pool-oriented" than a system-oriented measure since, for determining its score, it uses much more information about the pool than about the system under examination and this could be an explanation of its robustness to the pool reduction.

4. A MARKOVIAN USER MODEL

4.1 General framework

We will assume that each user starts from a chosen document in the ranked list and considers this document for a random time, that is distributed according to a known positive random variable. Then they decides, according to a probability law that we will specify in the sequel and independent from the random time spent in the first document, to move to another document in the list. Then, they considers this new document for a random time and moves, independently, to a third relevant document and so on.

After a random number of forward and backward movements along the ranked list, the user will end their search and we will evaluate the total utility provided by the system to them by taking the average of the precision of the judged relevant documents they has considered during their search. According to this construction when we compute this average, the precision of a document visited k times will contribute to the mean with a k/n weight.

We mathematically model the user behavior in the framework of the Markovian processes [23]. To fix the notation, we will denote by $X_0, X_1, X_2, \ldots$ the (random) sequence of document ranks visited by the user and by T_0, T_1, T_2 the random times spent, respectively, visiting the first document considered, the second one and so on. Therefore, $X_0 = i$ means that the user starts from the first document at rank i and $T_0 = t_0$ means that they spends t_0 units of time visiting this first document, then $X_1 = j$ means that they visits the document at rank j as the second one, and so on.

First of all, we will assume that X_0 is a random variable on $\mathcal{T} = \{1, 2, \ldots, T\}$ with a given distribution $\lambda = (\lambda_1, \ldots, \lambda_T)$; so for any $i \in \mathcal{T}$, $\mathbb{P}[X_0 = i] = \lambda_i$. Then, we will assume that the probability to pass from the document at rank i to the document at rank j will only depend on the starting rank i and not on the whole list of documents visited before.

This can be formalized as follows:

$$\begin{aligned} &\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] = \\ &= \mathbb{P}[X_{n+1} = j | X_n = i] = p_{i,j} \end{aligned} \quad (6)$$

for any $n \in \mathbb{N}$ and $i, j, i_0, \ldots, i_{n-1} \in \mathcal{T}$.

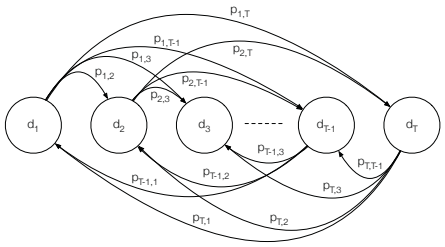

Figure 1: Structure of the Markov chain $(X_n)_{n \in \mathbb{N}}$.

Thanks to the condition (6) and fixing a starting distribution λ, the random variables $(X_n)_{n \in \mathbb{N}}$ define a time homogenous discrete time Markov Chain, shown in Figure 1, with state space $\mathcal{T}$, initial distribution λ and transition matrix $P = (p_{i,j})_{i,j \in \mathcal{T}}$ (Markov(λ,P) in the sequel).

To obtain a continuous-time Markov Chain, we have to assume that the holding times T_n have all exponential distribution, i.e.

$$\mathbb{P}[T_n \leq t] = \begin{cases} 0 & t < 0 \\ 1 - \exp(-\mu t) & t \geq 0 \end{cases}$$

Furthermore, conditioned on the fact that $X_n = i$, the law of T_n will be exponential with parameter μ_i, where μ_i is a positive real number that may depend on the specific state i of the chain the user is visiting at that time.

When our interest is only on the jump chain $(X_n)_{n \in \mathbb{N}}$, i.e. when we are interested in extracting the corresponding discrete-time Markov chain to act as a traditional evaluation measure, we simply assume that all these variables are exponential with parameter $\mu = 1$. When we are also interested in the time dimension, we have to provide a calibration for these exponential variables. We report a very simple example in Section 5 using click logs from Yandex.

The reason for choosing such a model will be immediately clear. Let us assume hereafter that the matrix P will be irreducible. This means that we can move in a finite number of steps from any document to any other document with positive probability. Thanks to (6) and the multiplication rule, the probability to pass in n steps from the document i to the document j is equal to $p_{i,j}^{(n)}$, the (i,j) entry of the matrix P^n and the irreducibility means that given any pair (i,j) there exists $n > 0$ such that $p_{i,j}^{(n)} > 0$. Furthermore, the probability distribution of any random variable X_n, which denotes the rank of the document visited after n movements, is completely determined by λ and P, since

$$\mathbb{P}[X_n = j] = (\lambda P^n)_j \ .$$

Given such a model, we assume that a user will visit a number n of documents in the list and then they will stop their search. In order to measure their satisfaction, we will evaluate the average of the precision of the ranks of the judged

relevant documents visited by the user during their search as

$$\frac{1}{n}\sum_{k=0}^{n-1}\text{Prec}(Y_k)\ .$$

where $(Y_n)_{n\in\mathbb{N}}$ denotes the sub-chain of $(X_n)_{n\in\mathbb{N}}$ that considers just the visits to the judged relevant documents at ranks $\mathcal{R}$, and shown in Figure 2.

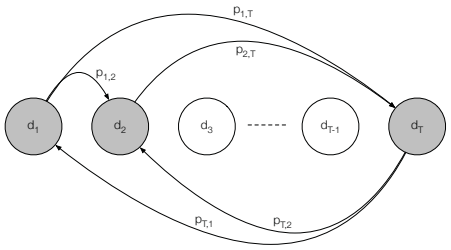

Figure 2: Structure of the sub-Markov chain $(Y_n)_{n\in\mathbb{N}}$ (relevant documents are shown in grey; not relevant ones in white).

Note that this sub-chain has in general a transition matrix different form P. The new transition matrix $\widetilde{P}$ can be computed easily from P by solving a linear system as detailed in [23] and discussed in Section 4.3.1. Note that $\widetilde{P}$ computed in this way somehow "absorbs" and takes into account also the probabilities of passing through not relevant documents (which are basically redistributed over the relevant ones) and makes it different from the transition matrix that you would have obtained by using only the relevant documents since the beginning.

Clearly the previous quantity is of little use if evaluated at an unknown finite step n. However, the Ergodic Theorem of the theory of the Markov processes is perfect for approximating this quantity:

THEOREM 1. *Let $\widetilde{P}$ be irreducible, λ be any distribution and $\mathcal{R}$ finite. If $(Y_n)_{n\geq 0}$ is Markov(λ,$\widetilde{P}$), then for any function $f:\mathcal{R}\to\mathbb{R}$ we have*

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=0}^{n-1}f(Y_k)\to\overline{f}\ \text{as}\ n\to\infty\right]=1$$

where $\overline{f}=\sum_{i\in\mathcal{R}}\pi_i f(i)$ and π is the invariant distribution of $\widetilde{P}$.

The importance of this class of theorems is clear: almost surely and independently of the initial distribution λ, we can approximate, for n large, the average over the time by the (much simpler) average over the states of the Markov chain. Indeed, under the previous assumptions it is possible to prove that the matrix $\widetilde{P}$ admits a unique invariant distribution, i.e a probability distribution π such that if $(Y_n)_{n\geq 0}$ is Markov(π,$\widetilde{P}$), then for any n

$$\mathbb{P}[Y_n=j]=\pi_j\ .$$

Moreover, the invariant distribution in this case is the unique left eigenvector of the eigenvalue 1 of the matrix $\widetilde{P}$, i.e. the unique solution of the linear equation

$$\pi=\pi\widetilde{P}\ .$$

REMARK 1. *Under additional hypotheses, it can be proved that the invariant distribution itself is the limit of any row of the matrix $\widetilde{P}^n$, as $n\to\infty$, useful result in order to evaluate in practice the invariant distribution. The convergence is generally very fast and for $n=10$ we already have a reasonable approximation of the true value of π. This justifies the use of MP to approximate the mean precision of the usually few documents visited by a user.*

We can now define a new family of user oriented retrieval effectiveness measures, called *Markov Precision (MP)*, which depends on the specific user model and the invariant distribution derived.

DEFINITION 1. *Given a ranked list of retrieved documents, defined by $\mathcal{R}$ the ranks of its judged relevant documents and defined a Markov (λ,P) user model, the Markov Precision metric will be defined as*

$$MP=\sum_{i\in\mathcal{R}}\pi_i Prec(i).$$

where $Prec(n)$ represent the Precision at n and π the (unique) invariant distribution of the Markov chain $(Y_n)_{n\in\mathbb{N}}$.

MP is defined without knowing the recall base RB of a given topic, but just the ranks of the judged relevant documents in a given run for this topic. As pointed out, for example in [22], the need to know the value of RB represents a weakness in AP that is overcome here.

In order to include the time dimension and thanks to the Ergodic Theorem for the continuous time Markov chains, we can replicate the previous computations and define a new measure

$$MPcont=\sum_{i\in\mathcal{R}}\widetilde{\pi}_i\text{Prec}(i).$$

where $\widetilde{\pi}_i=\frac{\pi_i(\mu_i)^{-1}}{\sum_{j\in\mathcal{R}}\pi_j(\mu_j)^{-1}}$, π denotes again the (unique) distribution of the Markov chain $(Y_n)_{n\in\mathbb{N}}$ and μ_i is the parameter of the holding time in state i. To use this alternative measure, we have to provide a calibration for the coefficients μ_i and we will compare MP with MPcont in a very simple example in Section 5 using click logs from Yandex.

4.2 Average Precision

In order to define a simple Markovian user model, whose MP value will be AP, let us consider the following transition probabilities among the documents in a given ranked list:

$$\mathbb{P}[X_{n+1}=j|X_n=i]=\tfrac{1}{T-1}\qquad(7)$$

for any $i,j\in\mathcal{T}$, $i\neq j$, and where, again, T denotes the cardinality of the set $\mathcal{T}$.

In this model we assume that a user moves from a document to another document with a fixed, constant probability, the value of which depends on the total number of relevant documents present in the specific run.

Since the invariant distribution is $\left(\frac{1}{T},\frac{1}{T},\ldots,\frac{1}{T}\right)$ we obtain that

$$MP=\frac{1}{T}\sum_{i\in\mathcal{R}}Prec(i)$$

which is equal to AP once multiplied by $\frac{T}{RB}$. Note that if we create the Markov chain starting directly from the relevant documents $\mathcal{R}$ we have to multiply MP by $Rec(T)$ as in

equation 3. In this way, we explain AP with a slightly richer user model, where the user can move forward and backward among any document and is not forced to visit only the relevant ones. It is also clear from the equation above that MP is not AP unless you provide it with the same amount of information AP knows about the recall base, namely rescaling MP by the recall base.

Looking at this the other way around, this instantiation of MP (without the rescaling) can be considered a kind of AP where the artificial knowledge of the recall base has been removed and so, it tells us how AP might look like if you remove the dependency on the recall base and insert an explicit user model. This consideration will turn out to be useful in the experimental part when we will find other user models, highly correlated to AP, which may give a richer explanation of it.

Moreover, the previous constant invariant distribution is common to many others user models. For example, if the transition matrix is irreducible and symmetric or even just bistochastic, meaning that the sum of the entries on each column is equal to 1, the invariant distribution is again the above constant vector. In this sense, if the validity of the present Markovian user model is accepted, it shows once more why AP has become a reference point, since it represents a good approximation for a wide class of models that we can define.

4.3 Other models

We will analyze three possible choices:

- *state space choice*: the Markov chain $(X_n)_{n\in\mathbb{N}}$ is on the whole set $\mathcal{T}$, indicated with `AD` (all documents model), or on the set $\mathcal{R}$, indicated with `OR` (only relevant documents model);
- *connectedness*: the nonzero transition probabilities are among all the documents, indicated with `GL` (global model), or only among adjacent documents, indicated with `LO` (local model);
- *transition probabilities*: the transition probabilities are proportional to the inverse of the distance, indicated with `ID` (inverse distance model), or to the inverse of the logarithm of the distance, indicated with `LID` (logarithmic inverse distance model).

We will obtain eight models that we will call after the possible three choices. So, for example, MP GL_AD_ID is an effectiveness measure with transition probabilities among all the retrieved documents, based on a model on the whole set $\mathcal{T}$, and with transition probabilities proportional to the inverse of the distance of the documents in the ranked list and so on for the other combinations of the parameters.

4.3.1 State space choice

In the `AD` case, we consider the whole Markov chain $(X_n)_{n\in\mathbb{N}}$ on the whole set $\mathcal{T}$ with a given initial distribution λ and a transition matrix $P = (p_{i,j})_{i,j\in\mathcal{T}}$ and then we derive the subchain $(Y_n)_{n\in\mathbb{N}}$ on the set $\mathcal{R}$. In order to obtain the invariant distribution of the subchain, we will have to derive its transition matrix $\widetilde{P}$. It can be proved (see [23]) that this matrix can be defined as follows

$$\widetilde{p}_{i,j} = h_i^j \quad \text{for} \quad i,j \in \mathcal{R}$$

Table 1: Main features of the adopted data sets.

	Topics	Runs	Min. Rel	Avg. Rel	Max. Rel
TREC 7	50	103	7	93.48	361
TREC 8	50	129	6	94.56	347
TREC 10	50	97	2	67.26	372
TREC 14	50	74	9	131.22	376

where the vector $(h_i^j, i \in \mathcal{T})$ is the minimal non-negative solution to the linear system

$$h_i^j = p_{i,j} + \sum_{k\neq\mathcal{R}} p_{ik} h_k^j \ . \tag{8}$$

So, once this linear system is solved, we obtain the transition matrix $\widetilde{P}$ needed to compute the Markov Precision for the given model.

In the `OR` model, we create the Markov Chain $(X_n)_{n\in\mathbb{N}}$ directly on the set $\mathcal{R}$.

4.3.2 Connectedness

In the `GL` model, we assume that the transition probabilities $p_{i,j} > 0$ for any choice of $i \neq j$. In this case we will assume that there will be a positive, even if very small, probability to pass from any document in the ranked list to any other. For example, the previous model for Average precision is a GL model

By contrast, in `LO` we will assume that there exist transition probabilities only among adjacent nodes. This is the same kind of logic behind RBP, even though RBP allows only for forward transitions, and is similar to the strategy of [8] for the two-dimensional placement problem.

4.3.3 Transition probabilities

In the `ID` model, we assume that the probability to pass from one document to another one in the ranked list is proportional to the inverse of the relative distance of these two documents:

$$\alpha(i,j) = \begin{cases} \frac{1}{|i-j|+1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{9}$$

Denoting by $(s_1, \ldots, s_m)$ the states of the Markov chain, we thus have the following transition probabilities:

$$p_{s_i,s_j} = \frac{\alpha(s_i,s_j)}{\sum_k \alpha(s_i,s_k)} \tag{10}$$

It is immediately clear that the probabilities (10) define an irreducible transition matrix P of a discrete time Markov Chain on the state space and therefore we can define Markov precision for this model.

In the `LID` model, we smooth the distance by using the base 10 logarithm so that that transition probabilities do not decrease not too fast. The choice of the base 10 for the logarithm is due to a typical Web scenario focused on the page of the first 10 results.

5. EVALUATION

5.1 Experimental Setup

In order to assess MP and compare it to the other pertinent evaluation measures (AP, P@10, Rprec, RBP, and

Table 2: Kendall τ correlation between AP and the other comparison measures using complete judgments (high correlations marked with *).

	AP	P@10	Rprec	bpref	RBP
TREC 7	1.000	0.8018	0.9261*	0.9275	0.7886
TREC 8	1.000	0.8264	0.9219*	0.9361*	0.8090
TREC 10	1.000	0.7551	0.8730	0.8896	0.7401
TREC 14	1.000	0.7295	0.9377	0.8394	0.7229

bpref), we conducted a correlation analysis and we studied its robustness to pool downsampling. As far as RBP is concerned, we set $p = 0.8$, which indicates a medium persistence of the user.

We used the following data sets: TREC 7 Ad Hoc, TREC 8 Ad Hoc, TREC 10 Web, and TREC 14 Robust, whose features are summarized in Table 1. We used all the topics and all the runs that retrieved at least one document per topic. In the case of collections with graded relevance assessment (TREC 10 and 14), we mapped them to binary relevance with a lenient strategy, i.e. both relevant and highly relevant documents have been mapped to relevant ones.

As far as pool downsampling is concerned, we used the same strategy of [4]: it basically creates separate random lists of relevant/not relevant documents and select a given fraction $R\%$ of them, ensuring that at least 1 relevant and 10 not relevant documents are in the pool. We used $R\% =$ [90, 70, 50, 30, 10].

As far as the calibration of time is concerned, we used click logs made available by Yandex [29] in the context of the Relevance Prediction Challenge[2]. The logs consist of 340,796,067 records with 30,717,251 unique queries, retrieving 10 URLs each. We used the training set where there are 5,191 assessed queries which correspond to 30,741,907 records and we selected those queries which appear at least in 100 sessions each to calibrate the time.

The full source code of the software used to conduct the experiments is available for download[3] in order to ease comparison and verification of the results.

5.2 Correlation Analysis

Table 2 reports the Kendall τ correlation [18] between AP and the other comparison measures, using complete judgements, for all the collections. Previous work [33, 34] considered correlations greater than 0.9 as equivalent rankings and correlations less than 0.8 as rankings containing noticeable differences. Table 2 is consistent with previous findings, with a high correlation between AP, Rprec, and bpref and lower correlation values for P@10 and RBP.

Table 3 reports the Kendall τ correlation between the different models for MP, discussed in Section 4.3 and whose notation (GL/LO, AD/OR, ID/LID) is used here as well, and the performance measures of direct comparison, for all the considered collections[4]. For each variant of MP, the table reports its actual value and also a second row labelled with the suffix @$Rec(T)$ to indicate a rescaled version of MP by recall. Indeed, this is the same operation needed to make MP equal to AP in the case of the model with constant transition probabilities discussed in Section 4.2 and corresponds to providing MP with the same level of information about the recall base that also AP uses. This has a twofold purpose: (i) to determine if there are other models beyond the ones of Section 4.2 which can give us an additional interpretation of AP; (ii) to get a general feeling of what is the impact of injecting information about the recall into an evaluation measure. In the table, we have marked high correlations, those above 0.90, with a star and we have marked extremely high correlations, those above 0.97, with two stars.

As a general trend MP tends not to have high correlations with the other evaluation measures, indicating that it takes a different angle from them. This can be accounted for by the effect of the user model explicitly embedded in MP which, for example, allows the user to move forward and backward in the result list while other measures allow only for sequential scans. On the other hand, the proposed models keep it not too far away from the other measures, especially those around precision (AP, P@10, Rprec), since the correlation never drops below 0.70. This is coherent with the fact that both MP and the other measures (AP, P@10, Rprec) are all around the concept of precision and so they have a common denominator.

Moreover, it can be noted that MP tends to be more correlated with P@10 and then with Rprec and AP. This is consistent with the fact that MP does not depend on the recall base, as P@10 does, while Rprec implicitly and AP explicitly depend on it.

Finally, the results show a moderate correlation with bpref and a slightly lower one with RBP, whose only common denominator is to not depend on the recall base.

Whit regard to @$Rec(T)$, we can note how they greatly boost the correlation with AP in almost all cases, often moving MP from low to high correlations, and, in turn, increase the correlation with Rprec and bpref (more correlated by themselves to AP) with respect to the one with RBP which tends to decrease.

In particular, there are some cases, like MP GL_AD_LID or MP LO_AD_ID, where it jumps between 0.97 and 1.00. We consider this a case in which MP is providing us with an alternative interpretation of AP, in the sense discussed in Section 4.2. For example, MP GL_AD_LID provided with information about recall tells us that we can look at AP as a measure that also models a user who can move backward and forward among all the documents in the list and who prefers smaller jumps to bigger ones. The fact that we have found a few models so highly correlated with AP suggests that AP has become a gold standard also because it represents some articulated user models.

5.3 Effect of Incompleteness on Absolute Performances

Figure 3 shows the effect of reducing the pool size on the absolute average performances, over all the topics and runs. For space reasons, we do not report figures for all the possible combinations reported in Table 3 but just some to give the reader an idea of the behavior of MP; the considerations made here are however valid also for the not reported figures.

It can be noted how MP shows consistent behavior over all the collections and for various models: its absolute aver-

[2] http://imat-relpred.yandex.ru/en/

[3] http://matters.dei.unipd.it/

[4] The fact that the values for the LO_AD_ID and LO_AD_LID models are the same is not due to a copy&paste error but to the fact that the two chains, in the local model, are the same apart from a constant and so they produce equal rankings.

Table 3: Kendall τ correlation between different instantiations of MP and the other comparison measures using complete judgments (high correlations marked with *; extremely high correlations marked with **).

	TREC 7					TREC 8				
	AP	P@10	Rprec	bpref	RBP	AP	P@10	Rprec	bpref	RBP
MP GL_AD_ID	0.7381	0.7522	0.7703	0.7827	0.7490	0.8997	0.8510	0.9074*	0.9222*	0.8382
MP GL_AD_ID@$Rec(T)$	0.9823**	0.7916	0.9243*	0.9322*	0.7799	0.9815**	0.8128	0.9217*	0.9299*	0.7938
MP GL_AD_LID	0.7378	0.7638	0.7712	0.7802	0.7632	0.8912	0.8641	0.9033*	0.9173*	0.8551
MP GL_AD_LID@$Rec(T)$	0.9954**	0.7994	0.9252*	0.9277*	0.7858	0.9953**	0.8221	0.9209*	0.9337*	0.8041
MP GL_OR_ID	0.7322	0.8311	0.7797	0.7689	0.7689	0.8162	0.9081*	0.8349	0.8402	0.9152*
MP GL_OR_ID@$Rec(T)$	0.9117*	0.8316	0.8937	0.8848	0.8243	0.9208*	0.8756	0.9024*	0.9145*	0.8637
MP GL_OR_LID	0.7379	0.7853	0.7782	0.7788	0.7858	0.8664	0.8884	0.8853	0.8947	0.8858
MP GL_OR_LID@$Rec(T)$	0.9726**	0.8158	0.9238*	0.9232*	0.8029	0.9722**	0.8477	0.9281*	0.9390*	0.8324
MP LO_AD_ID	0.7435	0.7706	0.7706	0.7874	0.7685	0.8931	0.8642	0.9011*	0.9174*	0.8537
MP LO_AD_ID@$Rec(T)$	0.9946**	0.7994	0.9225*	0.9265*	0.7858	0.9953**	0.8248	0.9219*	0.9343*	0.8066
MP LO_AD_ID	0.7435	0.7706	0.7706	0.7874	0.7685	0.8931	0.8642	0.9011*	0.9174*	0.8537
MP LO_AD_ID@$Rec(T)$	0.9946**	0.7994	0.9225*	0.9265*	0.7858	0.9953**	0.8248	0.9219*	0.9343*	0.8066
MP LO_OR_ID	0.7271	0.8229	0.7754	0.7634	0.8393	0.8138	0.9013*	0.8305	0.8354	0.9176*
MP LO_OR_ID@$Rec(T)$	0.9130*	0.8283	0.8958	0.8853	0.8211	0.9195*	0.9195*	0.8714	0.8987	0.9127*
MP LO_OR_LID	0.7386	0.8065	0.7826	0.7787	0.8058	0.8534	0.8982	0.8708	0.8810	0.8995
MP LO_OR_LID@$Rec(T)$	0.9552*	0.8278	0.9166*	0.9142*	0.8164	0.9506*	0.8623	0.9186*	0.9319*	0.8466

	TREC 10					TREC 14				
	AP	P@10	Rprec	bpref	RBP	AP	P@10	Rprec	bpref	RBP
MP GL_AD_ID	0.7264	0.7832	0.7727	0.7611	0.8013	0.8351	0.8078	0.8566	0.7778	0.7980
MP GL_AD_ID@$Rec(T)$	0.9726**	0.7340	0.8631	0.8771	0.8771	0.9896**	0.7221	0.9333*	0.8360	0.7140
MP GL_AD_LID	0.7125	0.7971	0.7633	0.7494	0.8187	0.8294	0.8185	0.8501	0.7751	0.8071
MP GL_AD_LID@$Rec(T)$	0.9941**	0.7512	0.8707	0.8878	0.7360	0.9977**	0.7303	0.9385	0.8397	0.8397
MP GL_OR_ID	0.7034	0.8269	0.7663	0.7470	0.8590	0.7968	0.8461	0.8206	0.7677	0.8302
MP GL_OR_ID@$Rec(T)$	0.9117*	0.8316	0.8937	0.8848	0.8243	0.9601*	0.7526	0.9327*	0.8650	0.7444
MP GL_OR_LID	0.7052	0.8077	0.7672	0.7466	0.8396	0.8140	0.8291	0.8348	0.7716	0.8155
MP GL_OR_LID@$Rec(T)$	0.9738**	0.7575	0.8740	0.8916	0.7448	0.9924**	0.7375	0.9398*	0.8432	0.7293
MP LO_AD_ID	0.7240	0.7969	0.7703	0.7614	0.8159	0.8297	0.8180	0.8504	0.7783	0.8089
MP LO_AD_ID@$Rec(T)$	0.9742**	0.7376	0.8654	0.8802	0.7218	0.9970**	0.7295	0.9363*	0.8405	0.7214
MP LO_AD_ID	0.7240	0.7969	0.7703	0.7614	0.8159	0.8297	0.8180	0.8504	0.7783	0.8089
MP LO_AD_ID@$Rec(T)$	0.9742**	0.7376	0.8654	0.8802	0.7218	0.9970**	0.7295	0.9363*	0.8405	0.7214
MP LO_OR_ID	0.7035	0.8300	0.7646	0.7449	0.8618	0.7997	0.8348	0.8234	0.7714	0.8220
MP LO_OR_ID@$Rec(T)$	0.9326*	0.7726	0.8767	0.8960	0.7618	0.9674*	0.7429*	0.9348*	0.8597	0.7377
MP LO_OR_LID	0.7114	0.8172	0.7676	0.7533	0.8472	0.8084	0.8324	0.8306	0.7689	0.8180
MP LO_OR_LID@$Rec(T)$	0.9579*	0.7601	0.8747	0.8949	0.7477	0.9877**	0.7372	0.9381*	0.8489	0.7306

age values decrease as the pool reduction rate increases in a manner similare to AP and Rprec. Consistently with previous results, P@10 and RBP exhibit a more marked decrease while bpref tends to stay constant. This positive property of bpref is an indicator that it is not very sensible or it does not fully exploit the additional information which is provided when the pool increases.

5.4 Effect of Incompleteness on Rank Correlation

Figure 4 shows the effect of reducing the pool size on the Kendall τ correlation between each measure on the full pool and the pool at a given reduction rate. The results shown are consistent with previous findings as far as the measures of direct comparison are concerned, showing that bpref is almost always the more robust measure to pool reduction. It is indeed plausible that, keeping bpref the absolute average performances almost constant, also the ranking of the systems does not change much.

As far as MP is concerned, we can note that global models [GL], shown in the case of TREC 7, 8 and 10, tend to perform comparably to AP and, when provided with the same information about the recall base, which both AP and bpref exploit, they consistently improve their performances and, in the case of TREC 8, they outperform AP and perform closely to bpref. This is an interesting result since, unlike bpref, the absolute average performances of MP vary at different pool reduction rates, indicating that MP is able to exploit the variable amount of information available at different pool reduction rates, still not affecting too much the overall ranking of the systems.

The global models [GL] on only relevant documents [OR] behave consistently with the global ones on all documents [AD], shown in the case of TREC 7 and TREC 10, even if they are a little bit more resilient to the pool reduction. This is consistent with the fact that they use less information than the AD ones and so they are less sensitive to the pool size. The TREC 7 also shows the effect of using the inverse of the distance [ID] or the log of the inverse of the distance [LID], which provides more robustness to pool reduction.

When it comes to local models [LO], these tend to behave comparably to the global ones in the case of all documents [AD], as can be noted in the case of TREC 8, while they are more affected by the pool reduction in the case of only relevant documents [OR], as can be noted in the case of TREC 14.

5.5 Time Calibration

On the basis of the click logs, 21% of the observed transitions are backward, a fact that validates our assumption that a user moves forward and backward along the ranked list.

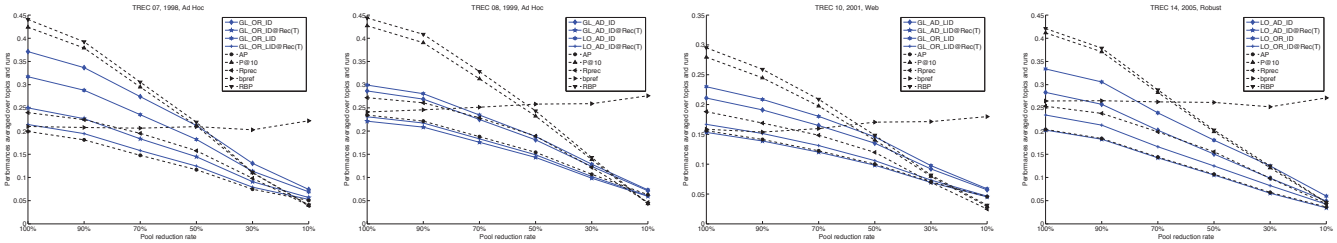

Figure 3: Pool reduction rate (x axis) vs. performance averaged over topics and runs (y axis)

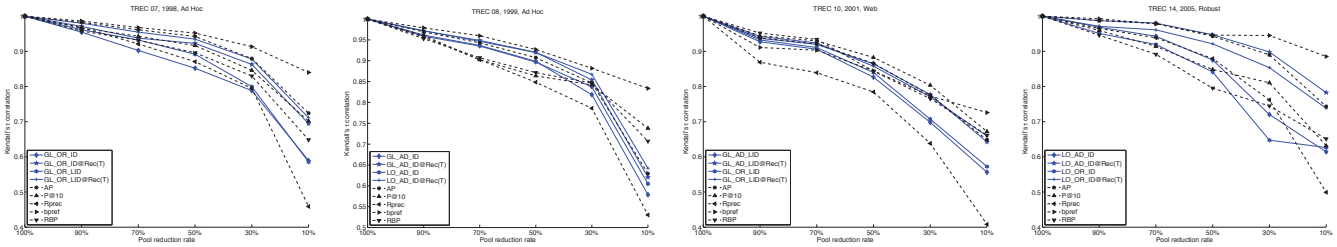

Figure 4: Pool reduction rate (x axis) vs. Kendall's rank correlation (y axis)

To compare the discrete-time version of MP with the continuous-time one, we have considered 3 runs with 5 relevant documents and estimated the parameters of the exponential holding times by the inverse of the sample mean of the time spent by the users visiting these states, multiplied by $(n-1)/n$. We used the GL_AD_ID model and the values of discrete-time MP and continuous-time MP are reported in Table 4.

Note that the precisions at each fixed rank n of the first, second and third runs are decreasing and as one expects MP of the three runs is decreasing. However, since the (estimated) holding times of the first documents in the first run are very low, continuos-time MP is smaller for the first run. This clearly shows that the use of continuous-time MP depends heavily on the calibration of the holding times.

6. CONCLUSIONS AND FUTURE WORK

We introduced a new family of measures, called MP, which exploit Markov chains in order to inject different user models and time into precision and which is not dependent on the recall base. This permitted us to overcome some of the traditional criticisms of AP (lack of a clear user model, dependence on the recall base) while still offering a measure which is AP when provided with the same amount of information about the recall base that AP exploits. Moreover, MP goes beyond almost all the evaluation measures allowing for non sequential scanning of the result lists.

We have proposed some basic user interaction models and validated their properties, in terms of correlation to other measures and robustness to pool reduction, thus showing it is as reliable as them. We have also found that some of these models have an extremely high correlation with AP and this can help in providing alternative interpretations of AP in the light of more complex user models and in explaining why AP is a "gold standard" in IR.

MP also bridges the gap between "rank-oriented" and "time-oriented" measures, providing a single unified framework where both viewpoints can co-exist and allowing for direct comparison among the values of the "rank-oriented" (discrete-time Markov chain) and "time-oriented" (continuous-time Markov chain) versions. We have also provided an example of how time can be calibrated using click logs from Yandex.

Future works concern the investigation of alternative user models able to account also for the number of relevant/not relevant documents visited so far – a kind of information which is actually available to a real user – by employing a multidimensional Markov chains to not violate the memoryless assumption. A further interesting option would also be to investigate whether click model-based IR measures [9] can be represented via the Markov chain and thus embedded in MP, i.e. whether the transition probabilities of the Markov chain can be learned directly from click-logs, thus leveraging models fully induced by user behaviour.

Another area of interest concerns how to calibrate time into MP: work on click model-based measures can shed some light in this respect and the techniques proposed by [30, 31] for calibrating time with respect to document length can link MP not only to click logs but also to document collections.

An interesting question for the future is whether MP could fit search tasks other than informational ones, such as fact, entity, or attributes focused searches or whether it could also work with other kinds of test collections, such as nugget-based ones [24].

Finally, the robustness of MP could be further investigated, for example evaluating how it performs on condensed-lists [27].

Acknowledgements

We wish to thank the anonymous reviewers and meta-reviewers whose comments and discussions helped us in improving the paper and better clarifying some angles of it.

The PREFORMA project[5] (contract no. 619568), as part of the 7th Framework Program of the European Commission, has partially supported the reported work.

[5] `http://www.preforma-project.eu/`

Table 4: Estimated parameters of the exponential holding times for three runs and values of the discrete-time and continuous-time MP.

Run	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	disc MP	cont MP
(1,1,1,1,0,0,0,1,0,0)	0.2000	0.0357	0.2000	0.0400	0.0056	0.0005	0.0035	0.0017	0.0034	0.0024	0.9205	0.6603
(1,1,1,0,1,0,0,0,1,0)	0.0177	0.0047	0.0037	0.0015	0.0041	0.0031	0.0057	0.0022	0.0061	0.0045	0.8668	0.8710
(1,1,0,1,1,0,0,0,0,1)	0.0056	0.0051	0.0062	0.0031	0.0046	0.0025	0.005	0.0022	0.007	0.005	0.8120	0.8001

7. REFERENCES

[1] J. A. Aslam, E. Yilmaz, and V. Pavlu. The Maximum Entropy Method for Analyzing Retrieval Measures. In SIGIR, pages 27–34, ACM, 2005.

[2] A. Broder. A Taxonomy of Web Search. *SIGIR Forum*, 36(2):3–10, 2002.

[3] C. Buckley and E. M. Voorhees. Evaluating Evaluation Measure Stability. In SIGIR, pages 33–40, ACM, 2000.

[4] C. Buckley and E. M. Voorhees. Retrieval Evaluation with Incomplete Information. In SIGIR, pages 25–32. ACM, 2004.

[5] C. Buckley and E. M. Voorhees. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, pages 53–78. MIT Press, USA, 2005.

[6] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In CIKM, pages 621–630. ACM, 2009.

[7] B. Carterette. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In SIGIR, pages 903–912. ACM, 2011.

[8] F. Chierichetti, R. Kumar, and P. Raghavan. Optimizing Two-Dimensional Search Results Presentation. In WSDM, pages 257–266, ACM, 2011.

[9] A. Chuklin, P. Serdyukov, and M. de Rijke. Click Model-Based Information Retrieval Metrics. In SIGIR, pages 493–502, ACM, 2013.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In WSDM, pages 84–75, ACM 2011.

[11] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., USA, 1997.

[12] K. Collins-Thompson and J. Callan. Query Expansion Using Random Walk Models. In CIKM, pages 704–711. ACM, 2005.

[13] C. Daniłowicz and J. Baliński. Document ranking based upon Markov chains. *IPM*, 37(4):623—637, July 2001.

[14] D. K. Harman. Overview of the Third Text REtrieval Conference (TREC-3). In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)* , pages 1–19. NIST, Special Pubblication 500-225, Washington, USA., 1994.

[15] D. K. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA, 2011.

[16] K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM TOIS*, 20(4):422–446, 2002.

[17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. In SIGIR, pages 154–161, ACM, 2005.

[18] M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945.

[19] J. Lafferty and C. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In SIGIR, pages 111–119, ACM, 2001.

[20] K. T. Maxwell and W. B. Croft. Compact Query Term Selection Using Topically Related Text. In SIGIR, pages 583–592, ACM 2013.

[21] A. Moffat, P. Thomas, and F. Scholer. Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In CIKM, pages 659–668. ACM, 2013.

[22] A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS*, 27(1):2:1–2:27, 2008.

[23] J. R. Norris. *Markov chains*. Cambridge University Press, UK, 1998.

[24] V. Pavlu, S. Rajput, P. B. Golbus, and J. A. Aslam. IR System Evaluation using Nugget-based Test Collections. In WSDM, pages 393–402, ACM, 2012.

[25] S. Robertson. A New Interpretation of Average Precision. In SIGIR, pages 689–690. ACM, 2008.

[26] T. Sakai. Ranking the NTCIR Systems Based on Multigrade Relevance. In AIRS 2004, pages 251–262. LNCS 3411, Springer, 2005.

[27] T. Sakai. Alternatives to Bpref. In SIGIR, pages 71–78, ACM, 2007.

[28] T. Sakai and Z. Dou. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In SIGIR, pages 473–482, ACM, 2013.

[29] P. Serdyukov, N. Craswell, and G. Dupret. WSCD2012: Workshop on Web Search Click Data 2012. In WSDM, pages 771–772. ACM, 2012.

[30] M. D. Smucker and C. L. A. Clarke. Stochastic Simulation of Time-Biased Gain. In CIKM, pages 2040–2044. ACM, 2012.

[31] M. D. Smucker and C. L. A. Clarke. Time-Based Calibration of Effectiveness Measures. In SIGIR, pages 95–104. ACM, 2012.

[32] I. Soboroff. Dynamic Test Collections: Measuring Search Effectiveness on the Live Web. In SIGIR, pages 276–283. ACM, 2006.

[33] E. Voorhees. Evaluation by Highly Relevant Documents. In SIGIR, pages 74–82, ACM, 2001.

[34] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM*, 36(5):697–716, 2000.

[35] E. Yilmaz and J. A. Aslam. Estimating Average Precision With Incomplete and Imperfect Judgments. In CIKM, pages 102–111. ACM, 2006.