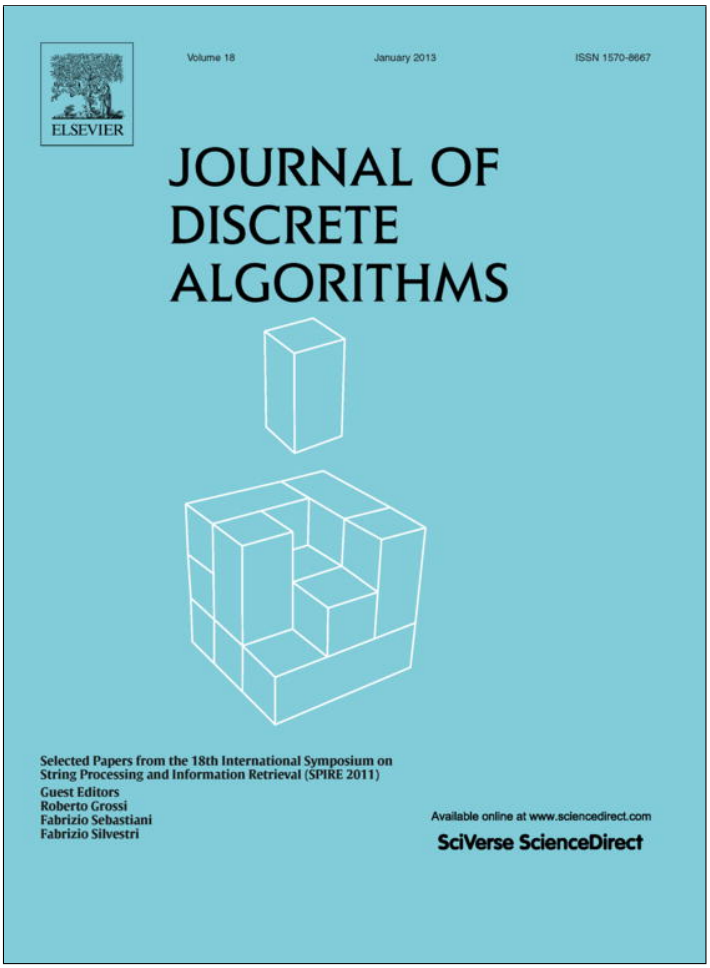

Journal of Discrete Algorithms 18 (2013) 49–62

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Model Based Comparison of Discounted Cumulative Gain and Average Precision

Georges Dupret [a,*], Benjamin Piwowarski [b]

[a] Yahoo! Labs, 4E4363, 701 First Avenue, Sunnyvale, CA, United States
[b] CNRS, University Paris 6, Paris, France

ARTICLE INFO

Article history:
Available online 13 November 2012

Keywords:
Discounted Cumulative Gain
DCG
Average Precision
User model
Web search
Metrics

ABSTRACT

In this paper, we propose to explain Discounted Cumulative Gain (DCG) as the expectation of the total utility collected by a user given a generative probabilistic model on how users browse the result page ranking list of a search engine. We contrast this with a generalization of Average Precision, pAP, that has been defined in Dupret and Piwowarski (2010) [13]. In both cases, user decision models coupled with Web search logs allow to estimate some parameters that are usually left to the designer of a metric. In this paper, we compare the user models for DCG and pAP at the interpretation and experimental level. DCG and AP are metrics computed before a ranking function is exposed to users and as such, their role is to *predict* the function performance. In counterpart to *prognostic* metric, a *diagnostic* metric is computed after observing the user interactions with the result list. A commonly used diagnostic metric is the clickthrough rate at position 1, for example. In this work we show that the same user model developed for DCG can be used to derive a *diagnostic* version of this metric. The same hold for pAP and any metric with a proper user model.

We show that not only does this diagnostic view provide new information, it also allows to define a new criterion for assessing a metric. In previous works based on user decision modeling, the performance of different metrics were compared indirectly in terms of the ability of the associated user model to predict future user actions. Here we propose a new and more direct criterion based on the ability of the *prognostic* version of the metric to predict the *diagnostic* performance.

0. Introduction

Optimizing the ranking of search engines, whether through the selection of ranking models, features, or the use of machine learning techniques, requires to accurately quantify the quality of document rankings. This in turn involves developing metrics that are robust (they quickly converge to their mean value when the number of queries increases), sensitive (they can order two search engines whose ranking is similar) and faithful (they measure user satisfaction). This paper focuses on the latter, and more precisely, on designing user models that explain the behavior observed through e.g. query search logs.

Following the work of Dupret et al. [11], the main argument of this paper is that deriving an accurate and reliable metric commands to define how users interact with a ranking list. Citing Robertson [23], "If we can interpret a measure (...) in terms of an explicit user model (...), this can only improve our understanding of what exactly the measure is measuring". More precisely, our view is that a metric is defined by two components:

* Corresponding author.
E-mail address: gdupret@yahoo-inc.com (G. Dupret).

1570-8667/$ – see front matter
http://dx.doi.org/10.1016/j.jda.2012.10.002

1. A user model that explains the behavior of the user as observed in search click logs.
2. A measure of performance which is defined based on the user model, as for example the expected utility of a user browsing the list of documents in the case of DCG.

This view resurfaced in the IR community the last years due to the (relative) availability of query search logs where parameters can be learnt. There is now an abundant literature on this topic [3–5,2,21,16,27,29].

Consequently, some of the parameters of a metric will be defined by the user model, and can thus be estimated from user interactions with search engines (i.e., search engine logs), while the others, related to the measure of performance itself, are left to the designer of the metric. One goal is to reduce this latter set to the minimum, in order to guide the design of a metric through the observed user behavior.

To further underline the importance of user models, let us consider the traditional 5 labels used to evaluate the Discounted Cumulative Gain or DCG. These characterize the relevance of a document to a query to be either PERFECT, EXCELLENT, GOOD, FAIR or BAD (P, E, G, F and B in short). Say a first ranking function – $F1$ – produces a sequence of documents with relevances $BBPBB$, while another function $F2$ produces $FFFBB$. Provided users scan the list sequentially, if one of them stops the search after the second position in the ranking, then he effectively sees BB if exposed to $F1$ and FF if exposed to $F2$. In this case, $F2$ is unambiguously better than $F1$. On the other hand, if a user scan at least three positions, then ranking $F1$ is arguably better. In conclusion, the user behavior defines which ranking is best.

Resorting to user modeling is also a first step to break the "chicken and egg" problem we face when comparing two different metrics: Deciding which metric is best calls for a third "meta" metric to compare the original metrics [9]. Because various "meta" metrics are likely to co-exist, a meta metric for the meta metrics is also necessary, etc.

Generative user models partly solve this problem because of their ability to predict which documents a user clicks when presented with a list of search results. By comparing the predicted clicks with the actual clicks observed on a (held out) set of sessions, we can identify which of several user models is best: if one model predicts more accurately future user interactions with a search engine than another, then the metric derived from the best user model is arguably better. This doesn't completely solve the problem though, as different metrics can be defined from a common user model.

Besides providing a more objective way of evaluating metrics, user models bring a valuable but under explored alternative to current metrics that compute an expected value of some utility given a user model. Contrasting with this "offline" view of measurement, the "online" view of metrics can be used to compare ranking functions *based on how users react to them* by using once again Web search logs. We will see that once a user model is defined, "online" metrics are in fact *diagnostic* version of a metric, which is defined as the expectation of the metric given an observed user behavior. This is to be contrasted to the *prognostic* version, where the metric, once its parameters are set, is computed without resorting to any search log. Diagnostic measures are interesting because they measure a performance which is closer to what is experienced by users of a search engine. In this paper, we illustrate these arguments based on the prognostic and diagnostic versions of DCG and AP in Section 5.1. We build upon the work presented by Dupret et al. in [13] and [10], with the following contributions:

1. Following a systematic presentation (likelihood, prognostic and diagnostic metric), we describe and analyze the DCG and AP user models in Sections 1 and 4, respectively. This allows to compare directly the two most used IR metrics and their possible user models.
2. We compare the user models using two criteria, (a) the likelihood of the observed logs in Section 5.1 and (b) the degree of matching between pro and diagnostic metrics in Section 5.3.

Notations and common assumptions. We first introduce some notations and common assumptions about the user behavior that we will use throughout the paper.

Because we suppose that all documents are judged, we can understand a ranking as a sequence of labels ℓ_r, $r = 1, \ldots, R$, where r indexes the position in the ranking. We often use the notation $\ell_{1:R}$ to represent the whole ranking up to position R. A user looking at a list of search results will only click on one of them if he previously actively looks at it (they are no "accidental" clicks). We say that a user always *examines* a result before clicking it and we define a binary variable E_r depending on the rank r, that indicates whether a particular rank r is examined by the user. The subscript r is dropped when there is no ambiguity. Finally, the binary variable C_r indicates whether a document was clicked or not.

We suppose that if a document is clicked, then its position is previously examined. We also use the following shorthand: e^+ and e^- are equivalent to "E is *true*" and "E is *false*", respectively. We also use $E = 1$ and $E = 0$ to denote e^+ and e^- when convenient. The same holds for c^+ and c^- or other binary variables introduced later. To shorten notations, we use a lowercase c as a shorthand for $C = c$ (and similarly for other random variables).

Finally, we will denote $\mathbf{s}$ a user session, as a shorthand to a series of clicks $c_{1:R}$ corresponding to one page of search results.

1. Discounted Cumulative Gain

Discounted Cumulative Gain (DCG) was proposed by Järvelin and Kekäläinen. It has the following general form:

$$\mathrm{DCG}_R = \sum_{1}^{R} D_r \mathrm{G}_r \tag{1}$$

where R is the maximum rank considered in the ranking, D_r is a discounting factor decreasing with the rank r and G_r is the gain achieved by presenting document d_r at rank r. Numerical values for the gains PERFECT, EXCELLENT, GOOD, FAIR and BAD often used in the literature are 10, 7, 3, 0.5 and 0. The discounting factor is typically set to be $D_r = 1/\log_2(1+r)$.

There are generally two related interpretations of this metric, originally an utilitarian one and more recently, with the development of user models in evaluation, a probabilistic one:

Utilitarian: The utility of a document to a user decreases when the document is low in the ranking.
Probabilistic: All documents are not examined with the same probability. This is motivated by the fact that search engine logs show that the probability of clicking a document decreases with its rank: It is natural to discount a document usefulness accordingly.

The original paper of Järvelin and Kekäläinen [18] introducing DCG doesn't relate it explicitly to user behavior or to a decision process, and is therefore closer to utilitarian interpretation. Departing from this, we propose here two user models (Sections 2 and 3) that lead to metrics interpretable as variant of DCG, each leading to a distinct set of discounting factor estimates. This work extends a previous work by Dupret [10].

If we postulate that a user who clicks on a document gains an amount of "utility" in relation with the document label ℓ, and if we postulate that the utility associated with a set of documents is simply the sum of their utilities, it seems natural to consider that a ranking is good if users who are presented with it achieve a large amount of utility. The corresponding metric is then:

$$\mathbb{E}(\text{total utility}) = \sum_{r=1}^{R} U(\ell_r) P(\text{click on document at position } r) \tag{2}$$

where $U(\ell_r)$ is the utility of the document at position r. This formula is a generalization of most user model based metrics, that has also been proposed in Moffat and Zobel [21], Yilmaz et al. [28] and Carterette [1] among others.

It is natural to associate the gain G_r to this utility, and the probability of a click to the discounting factor. We will see in the next sections that the situation is actually slightly more complex. Another observation is that in general the user decision to click depends on her previous actions, which in turn depends on the quality of the documents the user examined previously to reaching position r. We see that DCG is not able to cope with such a scenario because the discount factor depends exclusively on the rank. This is a limitation intrinsic to DCG that doesn't apply to Average Precision for example, as we will see in Section 4.

The metric in Eq. (2) is an expectation over all the possible sequences of clicks a user might choose to do. This is the *prognostic* version of the metric. Once a user actually clicked on a set of documents and ended her session, the utility she has gained is simply the sum of the utilities of the clicked documents. This sum is actually the *diagnostic* version of the metric. Of course, we expect the *prognostic* and the *diagnostic* versions to lead to values that are close to each other. In fact, the better we are able to predict the click – i.e. the better the user model, the closer they will be.

Given the definition of the metric, whatever the user model, the diagnostic metric will be the same. This is because the expected gain is in a close relationship with the actual clicks of the user. However, we will see (Sections 2.3 and 3.3) that the interpretation in terms of *gain* is substantially different depending on the user model.

In the following, we describe two possible user models for DCG, give the likelihood of a session (given the user model) and we derive an explicit form for prognostic and diagnostic versions of the metric.

2. Deterministic Click user model

The first user model is also the simplest: The user decides which rank to examine and click on the corresponding link. The gain is then the amount of utility associated with the document label.

User model 1 *(Deterministic Click).*

1. *The user chooses to examine a rank r between 1 and R with a probability $P(e_r^+)$.*
2. *She always clicks on the link to the document at rank r.*
3. *The document at rank r has a utility $\mathrm{U}(\ell)$ for her, where ℓ is the document label.*

This is a very crude and unrealistic model, but we will see that it casts doubt on the intuition that the discount factor at a given rank should be proportional to the probability of examining that rank.

This model is unable to predict sessions where no click occurs. Multiple clicks in a given session are understood as the same user repeating the above process[1] as many times as there are clicks in the session. In other words, a session with several clicks is really a sequence of one-click sessions.

Finally, the user model depends on one series of parameters, namely the probability $P(e_r^+)$ of examining the document at rank r. These probabilities can be learned from user search sessions by maximizing the likelihood defined next.

2.1. Likelihood

If search logs are available, then an estimate of $P(e_r^+)$ can be obtained by maximizing the likelihood. The likelihood of a single session with a click at position r is:

$$\mathcal{L}_r = P\left(e_r^+\right) \times \prod_{s=1;\, s\neq r}^{R} \left(1 - P\left(e_s^+\right)\right)$$

For a session $\mathbf{s}$ composed of c clicks at positions $r_1, \ldots, r_c$ (positions can be repeated), the likelihood is, as a consequence of the user model assumptions:

$$\mathcal{L}(\mathbf{s}) = \prod_{i=1}^{c} P\left(e_{r_i}^+\right) \times \prod_{s=1;\, s\neq r_i}^{R} \left(1 - P\left(e_s^+\right)\right)$$

The likelihood of a session can then be maximized in order to learn the model parameters, namely the probability $P(e_r^+)$. This is straightforward since we have no hidden variables, i.e. examination is equivalent to click. Hence, the probability of examining rank r is simply the ratio of the number of clicks on a document at rank r to the total number of clicks.

2.2. Prognostic metric

A reasonable measure of the ranking quality is the expected utility a user achieves given this ranking. We can easily compute this metric based on the current user model:

$$\begin{aligned}\mathbb{E}(\text{total utility}) &= \sum_{r=1}^{R} \mathtt{U}(\ell_r) P\left(c_r^+\right) \\ &= \sum_{r=1}^{R} \mathtt{U}(\ell_r) P\left(e_r^+\right)\end{aligned} \tag{3}$$

We see that it is possible to match this expected utility with the DCG if we make the following associations:

$$\begin{cases} \mathtt{G}_\ell = \mathtt{U}(\ell) \\ D_r = P\left(e_r^+\right) \end{cases} \tag{4}$$

This shows that if we assume that the Deterministic Click model holds, then the expected utility coincides with DCG. The converse is not true as we will illustrate later by introducing another model from which DCG can also be defined.

Stating that the discount factors are actually the probabilities of examination is tantamount to declaring that users behave as described by the Deterministic Click model. In the case of Web search, these assumptions are clearly unrealistic.

2.3. Diagnostic metric

We might want to measure the quality of the user experience *after* she finished her search. In this case we know which documents she clicked and we have:

$$\mathbb{E}(\text{total utility}|\mathbf{s}) = \sum_{r=1}^{R} U(\ell_r) c_r \tag{5}$$

where $c_r = 1$ if the user clicked at position r and 0 otherwise.

Because the utility of a document corresponds to the gain associated with its label as identified in Eq. (4), the diagnostic DCG can also be written:

$$\mathbb{E}(\text{total utility}|\mathbf{s}) = \sum_{r=1}^{R} \mathtt{G}_r c_r$$

This is simply the sum of the gains of the clicked documents.

[1] Nothing prevents clicking twice on the same document.

3. Probabilistic Click user model

The Deterministic Click model is unrealistic because, among other reasons, it assumes that the user chooses one position in the ranking and ignores the other. Once they choose a position, it is assumed that users

1. blindly click on the corresponding document. This is unrealistic because users don't click deterministically on a document they examine. Instead, they evaluate the document snippet before deciding whether to click;
2. completely ignore the snippets situated higher and lower in the ranking. Here again, users behave otherwise. Eye tracking experiments suggest they tend to browse the result list sequentially as shown by Granka et al. [14].

These observations suggest the following model.

User model 2 *(Probabilistic user model).*

1. *A user examines sequentially the ranking up to a position r chosen before starting to search. After reaching this position she abandons the search.*
2. *A user clicks on an examined document with a probability $P(c^+|e^+;\ell)$ where ℓ is the editorial label of the document.*

Item (2) requires that the search engine generates snippets that represent fairly the document content. In other word, we assume that the perceived and actual relevances match. If we had editorial labels for the document snippets as well as the documents themselves we could estimate the probability of a click given the snippet label rather than the document label. See the work of Turpin et al. [25] for a more complete discussion on the influence of snippet quality and the consequences of its discrepancy with document relevance.

We make no assumption on whether the user is satisfied or not when abandoning the search. The reason is related to the original definition of the DCG. If a user met her information need, i.e. she is satisfied, she will abandon the search. Hence her decision to examine a particular position will depend on the document she has seen previously. But the discounting factors appearing within the DCG definition depend exclusively on the position so if we admit a dependency between examination and satisfaction, we will not be able to factorize $\mathbb{E}$(total utility) into gains and discounting factors.

3.1. Likelihood

We first define for convenience the multinomial variable A on $\{1,\dots,R\}$ that describes the position up to which the user is willing to examine the ranking: $A=r$ means that she decides to end the search at position r. Note that $A=r$ entails $\{e^+_{1:r}, e^-_{r+1:R}\}$, i.e. all snippets up to position r are examined and none is examined after r.

The user model is completely specified by the following parameters:

- The probability $P(A=r)$ that the user decides to examine the first r positions. $A=r$ entails $e^+_{r'}$ for any $r'\leqslant r$ and $e^-_{r'}$ for $r'>r$.
- The probability of clicking on a document when examined, $P(c^+|e^+;\ell)$, as a function of the document label.

Because we have defined a user model, we are able to predict which documents a user will click during a session and hence estimate the model parameters by maximum likelihood. The likelihood is obtained by marginalizing out the (partially) hidden variables A and $E_{1:R}$.

As shown in Appendix A, the likelihood of a session $\mathbf{s}$ with a last click on position b can be written:

$$\mathcal{L}(\mathbf{s}) = \prod_{r=1}^{b} P\left(c_r\middle|e_r^+;\ell_r\right) \times \sum_{a=b}^{R} P(a) \prod_{r:\ a\geqslant r>b} P\left(c_r\middle|e_r^+;\ell_r\right) \tag{6}$$

where the first factor corresponds to the likelihood of all the clicks up to the last click and the second corresponds to the likelihood from after the last click, where we have to sum over all the remaining possible values for A.

It is a simple matter to multiply the likelihood of a set of observed sessions and maximize it with respect to $P(c_r^+|e_r^+;\ell_r)$ and $P(A=r)$ to obtain estimates of these probabilities. We used the Expectation Maximization algorithm [7] for this task.

Note that the utilities U_ℓ do not appear in Eq. (6) and hence cannot be estimated by maximum likelihood.

3.2. Prognostic metric

The expected utility can be expressed as (Appendix A):

$$\mathbb{E}(\text{total utility}) = \sum_{r=1}^{R} \mathrm{U}(\ell_r) P\left(c_r^+\middle|e_r^+;\ell_r\right) P(A\geqslant r) \tag{7}$$

where the terms $\mathtt{U}(\ell_r)P(c_r^+|e_r^+;\ell_r)$ only depend on the document label, not the rank,[2] and can therefore be identified with the DCG gain $\mathtt{G}_r$ in Eq. (1). The term $P(A \geqslant r)$ only depends on the position in the ranking and is associated with the discounting factor[3]:

$$\begin{cases} \mathtt{G}_\ell = \mathtt{U}(\ell) P\left(c^+ \middle| e^+; \ell\right) \\ D_r \propto P(A \geqslant r) \end{cases} \tag{8}$$

The gain can therefore be interpreted as the *expected* utility and not the utility itself, the distinction arising because documents are not always clicked. Another way to state this is to recognize that the utility gained by a user who clicks on a document with label ℓ is now $\mathtt{G}_\ell / P(c^+|e^+;\ell)$ instead of $\mathtt{G}_\ell$ as predicted by the first model.

3.3. Diagnostic metric

As for the Deterministic Click model, the diagnostic version of the metric is the sum of the utility of the documents the user has clicked:

$$\mathbb{E}(\text{total utility}|\mathbf{s}) = \sum_{r=1}^{R} \mathtt{U}(\ell_r) c_r$$

However, we have seen in Eq. (8) that the utility is not directly equal to the gain as in the Deterministic Click model. In terms of the gains, the diagnostic version of the metric can be written:

$$\mathbb{E}(\text{total utility}|C_{1:R}) = \sum_{r=1}^{R} \frac{\mathtt{G}_r}{P(c^+|e^+;\ell_r)} c_r$$

The gain associated with a document is simply rescaled to take into account its probability of being clicked. This is to be contrasted to the diagnostic version of total utility with the deterministic user model in Eq. (5), where the gain is simply the utility of the corresponding document. From a theoretical point of view, this is the consequence of the fact that the user can examine a rank *without* clicking on the document.

Moffat and Zobel [21] proposed the RBP metric as an improvement over Average Precision. As it turns out RBP is similar to DCG with numerical values for the discount factor based on $P(e_{r+1}^+|e_r^+) = p$ where p is adjusted to the click data. The main difference in this work is that $P(e_{r+1}^+|e_r^+)$ is a set of free parameters – one for each rank r, while it is parameterized by the "patience" parameter in RBP. Carterette [1] also reviews a whole series of possible parameterizations.

4. Probabilistic Average Precision

Unlike DCG that accommodates multi-grade editorial assessment, Average Precision [26] assumes that documents are either relevant or irrelevant.[4] It is defined as the average of the precisions computed at the relevant document positions:

$$\mathtt{AP} = \frac{1}{T} \sum_{r=1}^{\infty} \text{precision at } r \times \text{ relevance at } r \tag{9}$$

where T is the number of documents relevant to the query at hand and "relevance at r" is 1 if the document is relevant and 0 otherwise. In practice, the sum is often truncated. In order to compare the AP and DCG user models more easily, let us denote R the number of ranks considered.

DCG and AP are the two most popular Information Retrieval metrics, with DCG more common in Web search applications while AP is used more often in "pure" Information Retrieval tasks. A reason often cited is that DCG favors high precision, which is more important for Web search as typically many different documents potentially satisfy an information need. In "pure" Information Retrieval a greater emphasis is put on retrieving as many relevant document as possible, i.e. the recall performance is more important. AP is more adequate in this case because as shown in Eq. (9) all relevant documents participate in the evaluation.

Like DCG, the Average Precision AP [26] metric can be associated with a particular set of hypotheses on the user behavior. In [13], the authors propose the following model:

[2] The presence of r in the expression $\mathtt{U}_r P(c_r^+|e_r^+;\ell_r)$ doesn't imply a dependence on r of the gains; it's role is to identify the document label. For example, $\mathtt{U}_r P(c_r^+|e_r^+;\ell_r) = \mathtt{U}_s P(c_s^+|e_s^+;\ell_s)$ as long as the documents at positions r and s share the same label.

[3] It is always possible to derive discounting factors $D_{1:R}$ from the probabilities $P(A \geqslant r)$, $r = 1, \ldots, R$, but the opposite is not true because D_r is not required to be a probability in the original definition [18].

[4] This metric has been adapted by Robertson et al. [24] to multi-grade assessments but this does not change fundamentally the user model, see Section 6.

User model 3 *(Probabilistic* AP*).* [5]

1. *The user decides before hand the number $N = n$ of relevant documents she needs to meet her information need.*
2. *She browses the result list sequentially.*
3. *She clicks on a document she examines with a probability $P(c_r|e_r;\ell_r)$ that depends on the binary document label: $\ell_r = \ell^-$ or ℓ^+ depending on whether the document is relevant or not, respectively.*
4. *She ends her search as soon as she clicked on enough relevant documents to satisfy her information need.*

Because different users need a different number of documents, N is a discrete random variable with distribution $P(N)$. We see that this model assumes that a user ends her search only if she is "satisfied" and that a search must end on a relevant document.

This user model in fact leads to a generalization of AP. To recover the exact AP metric, some additional assumptions need to be made:

User model 4 *(Additional assumptions for* AP*).*

1. *If there are T relevant documents for a query, the probability that a user needs exactly n relevant documents to satisfy her information need is uniform, i.e. $P(N = n) = \frac{1}{T}$.*
2. *A user always clicks on a relevant document she examines: $P(c_r^+|e_r^+;\ell_r^+) = 1$.*

These additional hypotheses are strong, especially the first one. Moreover, the reliance of AP on the prior knowledge of the number of relevant documents has been often cited as a limitation [21]. The pAP model relaxes these hypotheses and makes it possible to evaluate both $P(N = n)$ and $P(c_r|e_r^+;\ell_r^+)$ from the clickthrough logs.

It is interesting to compare the DCG and pAP model hypotheses: Both models suppose that users browse the results sequentially, but the stopping criterion is different. The DCG user model supposes that a user pre-defines the number of ranks she is willing to browse. In some sense, the effort – evaluated as a number of ranks – the user is ready to devote to the search is fixed, limited and independent of the results she finds. The DCG metric is then simply the expected total amount of "gain" the user achieves given the effort.

On the other hand, a pAP user starts her search by pre-defining a number of relevant documents she requires and she ends her search only when she finds that many documents, independently of the amount of effort this might involve. If we consider that the required number of documents she wants to retrieve is a measure of the "gain" she wants to achieve, then pAP can be understood as a measure of the effort the user need to invest in order to achieve a certain amount of "gain".

Informally, DCG fixes the "effort" and uses the "gain"as the metric, while for pAP fixes the "gain" and uses the "effort". This also holds for AP as the only difference between these two metrics are the numerical values given to $P(N = n)$ and $P(c_r|e_r^+;\ell_r^+)$. The distinction between "effort" and "gain" metric was noted earlier by Dupret [9] and was taken on again in [1] by Carterette.

The likelihood, prognostic and diagnostic measures, as reported in [13], are described below.

4.1. Likelihood

The likelihood of a session **s** is defined as

$$\mathcal{L}(\mathbf{s}) = P(n_b)\prod_{r=1}^{b} P\left(c_r\middle|e_r^+;\ell_r\right) + P(N > n_b)\prod_{r=1}^{R} P\left(c_r\middle|e_r^+;\ell_r\right)$$

where n_b is the number of relevant documents clicked within the session **s**. We can see that the likelihood is decomposed in two terms, the first corresponding to the case where the user information need was satisfied, and the second where it was not (the user clicked less relevant documents that she wanted to see).

We observe that the likelihood for the AP user model is formally similar to the likelihood for the probabilistic DCG user model – Eq. (6), the difference being the stopping criterion that divides the list of documents into the examined and the non-examined ones.

4.2. Prognostic metric

The expected value of the AP given a ranking $\ell_{1:R}$ is:

[5] pAP for short.

Table 1
Probabilistic Click model: Median probability of click given a label, DCG gains and utilities.

Label	B	F	G	E	P
$P(c^+\|e^+;\ell)$	0.27	0.27	0.34	0.37	0.85
$\mathrm{U}(\ell)$	0.00	1.85	8.82	18.92	11.76
$\mathrm{G}(\ell)$	0.00	0.50	3.00	7.00	10.00

$$\mathbb{E}(\mathrm{AP}) = \sum_{n\geqslant 1} P(n) \sum_{r=1}^{R} \delta_{\ell_r^+} \mu_+ \frac{n}{r} \underbrace{\binom{t_{r-1}}{n-1} \mu_+^{n-1} (1-\mu_+)^{t_{r-1}-n+1}}_{(*)}$$

where t_r is the number of relevant document up to rank r, $\delta_{\ell_r^+}$ is 1 if the document at rank r is relevant and 0 otherwise and μ_+ is the probability of clicking on a document given that it is examined and relevant. The term $(*)$ corresponds to the probability of clicking on $n-1$ relevant documents among t_{r-1}.

4.3. Diagnostic metric

Finally, the diagnostic version of AP is given by:

$$\mathbb{E}(\mathrm{AP}|\mathbf{s}) = \delta_{\ell_b^+} \times \frac{P(n_b)}{P(n_b) + P(N > n_b) \prod_{r=b+1}^{R} P(c_r^-|e_r^+;\ell_r)} \times \frac{n_b}{b}$$

which is composed of three terms that respectively (1) set the value to 0 if the user did her last click on a non-relevant document (hence the user stopped before finding n documents), (2) the probability that the user wanted to see n_b relevant document and (3) the precision at rank b.

Compared to the DCG user model, the AP user model is somehow more complex because the stopping criterion depends on the list of document labels. It corresponds to the precision achieved at the last clicked rank if the user wanted to see exactly n_b relevant documents.

5. Comparing user models

In this section, we compare the different user models and their associated metrics from two points of view:

1. The likelihood of the model (Section 5.1). This is the method used in [11] and subsequently used by many authors [16, 15,30,17].
2. The correlation between diagnostic and prognostic (Section 5.3).

The experiments were conducted on logs of a commercial search engine corresponding to a set of approximately 33 000 unique sessions for which we had an editorial judgment on a 5 level scales for each of the top 10 urls, together with a record of which urls had been clicked. Each record in our data set has the following form: A sequence of 10 labels $\ell_{1:10}$ followed by a sequence of 10 *True* or *False* tokens that indicates the states of $C_{1:10}$. The approximately 300 queries present in the logs have been selected randomly among the set of unique queries having at least 10 sessions, i.e. the probability of picking a given query is independent of the frequency of the query in the logs. The number of sessions of a given query is capped at 500 to limit the influence of highly frequent queries.

5.1. Likelihood of user models

5.1.1. Comparing *DCG* models

The first DCG model is unable to predict sessions without clicks so we had to remove those sessions. In order to compare the performance of the two DCG models, i.e. the Deterministic and the Probabilistic Click models, we need to train and test these models on the same data, so we trained the Probabilistic Click model on the set where the sessions have at least one click. These results are reported first.

We divided the data in 10 random subsets in such a way that all sessions of a query fall into the same subset. We use each of these subsets (i.e. 10% of the original set) as the data we maximize the likelihood on.

This results in 10 different sets of estimates for the parameters of User models 1 and 2 that we report in Table 1, first row and in Table 2. We also included in this last table for comparison a set of discounting factors that are often chosen by default.

Figs. 1 and 2 present the same results graphically. The boxplots attest that the estimates turn out to be fairly stable. The boxplot in Fig. 1 reports the probability of examination of each rank for the Deterministic Click model. Results agree with intuition.

Table 2
Left column: Mean probabilities of examination $P(e_r^+)$ for the Deterministic Click model. Right columns: The probability of abandoning beyond rank r for the Probabilistic Click model and, for comparison, the popular $D_r = 1/\log_2(1+r)$ discount factors for $r = 1, \ldots, 10$. We observe that the empirical discounting factors decrease faster beyond rank 3 than what the logarithmic decay accounts for.

Rank r	$P(e_r^+)$	$P(A \geqslant r)$	$1/\log_2(1+r)$
1	0.53	1.00	1.00
2	0.16	0.70	0.63
3	0.10	0.47	0.50
4	0.06	0.32	0.43
5	0.04	0.23	0.39
6	0.03	0.17	0.36
7	0.03	0.13	0.33
8	0.02	0.09	0.32
9	0.02	0.07	0.30
10	0.01	0.05	0.29

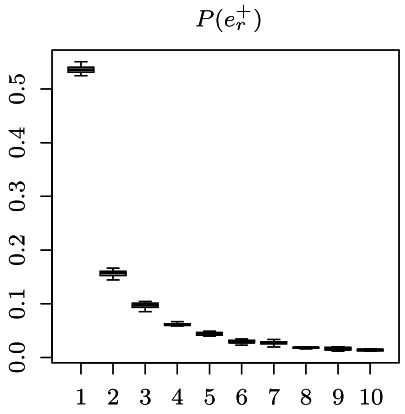

Fig. 1. Probabilities $P(e^+)$ for the Deterministic Click model.

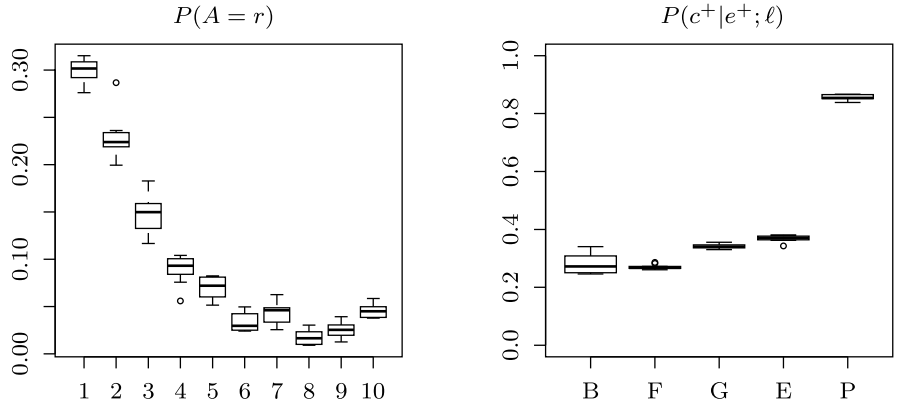

Fig. 2. Probability $P(A = r)$ of abandoning the search at rank r and probability $P(c^+|e^+, \ell)$ of click given a label for the probabilistic DCG user model.

In Fig. 2, the left boxplot reports the probability of abandoning the search at rank r for the probabilistic user model. Because abandoning at rank r entails in this model that the user examined all ranks up to r, these probabilities cannot be compared easily with the probabilities derived from the deterministic user model. Note that more users decide to end their search at ranks 9 or 10 than 8. Intuitively there is no contradiction because a user who examines up to rank 8 might as well go until the end of the page. The boxplot on the right reports the probability of click for examined links for the 5 different labels. Although nothing in the model enforces it, we see that the model predicts that documents with a better label also have a higher probability of being clicked. BAD and FAIR documents have very similar probability of being clicked, as do GOOD and EXCELLENT. PERFECT stands out as a category of document with a particularly high probability of being clicked. This makes sense as this label is used for pages that are the target of a navigational query.

We have proposed two distinct models to explain DCG, the Deterministic and the Probabilistic Click user models. The second model seems more realistic, but we would like to confirm quantitatively this intuition. Both models are generative models and can be used to predict user behavior; we can therefore compare the accuracy of these predictions on the test sets. We use the *perplexity* [8] – a common measure of the "surprise" of a model when presented with a new observation.

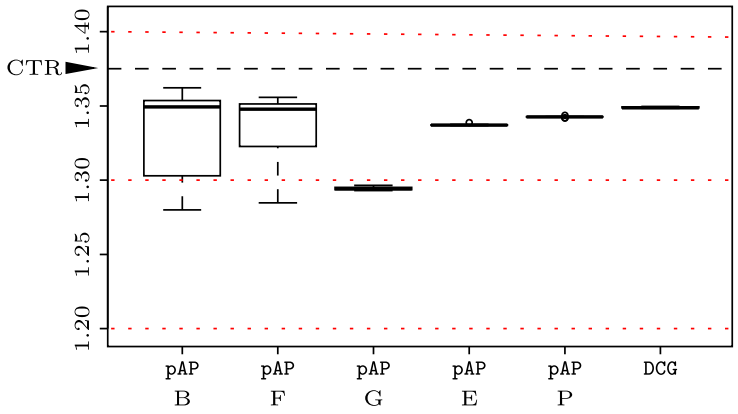

Fig. 3. Perplexity of the Probabilistic Click user model (DCG) and the Probabilistic Average Precision model (pAP). The label below "AP" represents the threshold at which a document is considered as relevant. Boxplots are based on 10 folds cross-validation.

(To the best of our knowledge, [12] were the first to use perplexity as way to compare user models.) Given a proposed probability model q of the true distribution p, one may evaluate q by asking how well it predicts a separate test sample of size N also drawn from p. The perplexity of the model q is defined as

$$2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i)}$$

where $q(x_i)$ is the probability of event or observation x_i according to model q. Better models q of the unknown distribution p will tend to assign higher probabilities to the test events. Thus, they have lower perplexity: they are less surprised by the test sample.

In the context of the user behaviors, the perplexity is a monotonically increasing function of the joint probability of the sessions in the test set. Analytically, this probability is identical to the likelihood of the test set, but instead of maximizing it with respect to the parameters, those are held fixed at the values that maximize the likelihood on the *training* set.

All the sessions in both the training and test sets have exactly 10 results per page ($R = 10$) so that by setting N to 10 times the number of sessions, the perplexity can be loosely[6] interpreted as the number of trials per correct prediction of a binary event: the click or skip of a document. The lower the perplexity, the better the model: A perplexity of 1 corresponds to perfect prediction, while a perplexity of 2 corresponds to randomly predicting the two possible outcomes with 50% chances. Perplexity larger than two characterizes models that are so bad that simply inverting the binary predictions would lead to a more accurate model.

The mean perplexity of the Deterministic Click model evaluated on the 10 random splits of the data is 1.29, while the mean perplexity of the Probabilistic Click model is 1.27. A Welch Two Sample t-test lead to a p-value smaller than 2.2e−16. The Probabilistic Click model is therefore statistically significantly better at predicting the user behavior.

5.1.2. Comparing DCG and AP

To compare the Probabilistic Click model and the Probabilistic Average Precision models, it is more realistic to include the sessions without clicks. We follow the same procedure as above regarding cross-validation. The pAP model can only accommodate binary labels. We therefore tried the 5 different possible mappings: Mapping G in Fig. 3 for example means that only the documents with a label equal or superior to GOOD are considered relevant. The B mapping is the special case where all documents are considered as relevant.

It stands out that the best threshold to decide whether a document is relevant is GOOD for the pAP model. The BAD and FAIR thresholds occasionally lead to low perplexities, but the performance is very variable and the median lays around 1.35, which is significantly worse than the score achieved by selecting GOOD as the threshold. More importantly, we see that the Probabilistic Click model (DCG) performance is significantly worse than the best pAP model. This suggests that for the particular data set we have, evidences are that the user behavior is best represented by the Probabilistic Average Precision model and that as a consequence, evaluation based on the pAP metric is more adequate. This comes as a surprise because the data we used comes from Web search and DCG is usually considered more adequate for this task than AP. On the other hand, pAP differs in some important respects from AP and benefits from being more flexible.

5.2. Gains

Although none of the DCG models provides a method to evaluate the gains, we can use the probability of clicks given an editorial label to estimate the utilities. Returning to Eq. (8) we see that we have the relation

[6] This interpretation is not strictly correct because the clicks and skips in a session are not independent. The evaluation itself continues however to be valid.

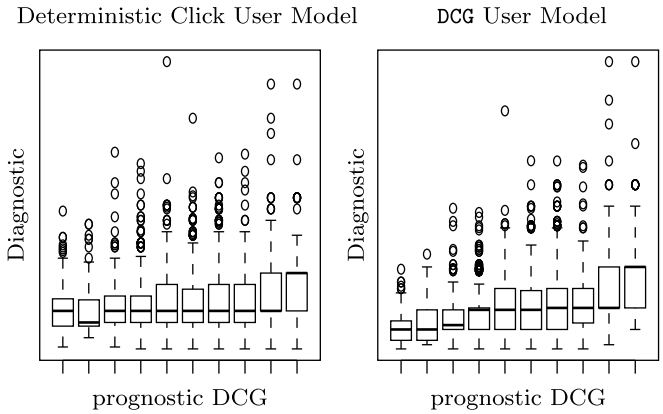

Fig. 4. Diagnostic vs. prognostic DCG for the Deterministic Click and the DCG user models.

$$\mathrm{U}(\ell_r) = P\left(c_r^+ \middle| e_r^+; \ell_r\right)/\mathrm{G}_r$$

between the DCG gains and the utilities defined in the Probabilistic Click model. If we plug-in the gain values commonly used into this formula, we obtain the results reported in Table 1, second and third rows. This seems to indicate that the gains associated with EXCELLENT and PERFECT values are inadequate because they don't respect the order of the labels on the utility scale. Note also that in order to evaluate the diagnostic DCG we must use the utility, not the gain. The sum of the gains of the clicked documents doesn't lead to the correct estimate unless utility and gains happen to be equal like in the Deterministic Click model.

5.3. Prognostic vs. diagnostic

We can qualify the DCG as a *prognostic* metric because it is typically computed to evaluate a ranking function *before* it is presented to users. Its aim is to predict whether a new ranking function is more likely to satisfy users. Once sufficient data is collected on the interactions of users with the new ranking function, a *diagnostic* metric can be evaluated: Instead of computing the expected utility of the ranking, we compute its average utility knowing the set of sessions, i.e. as the average of the sum of the utilities of the clicked documents. For example, for one session, the *diagnostic* DCG of a session $\{c_1^-, c_2^+, c_3^-, c_4^+, c_{5:10}^-\}$ is $\hat{\mathrm{U}} = \mathrm{U}(\ell_2) + \mathrm{U}(\ell_4)$. Note that it is not possible to associate a diagnostic metric to a prognostic metric like DCG unless a user model is defined. The reason is that different user model can lead to different diagnostic metrics even when the prognostic metric is the same as we have illustrated with the two DCG models.

Moreover, the correlation between a diagnostic and prognostic version of a metric gives a further indication of the goodness of the user model. In theory, if the user model corresponds to what is actually the behavior of users, the *diagnostic metric should converge towards the prognostic metric* as the number of observed sessions increases.

This brings new criteria to the evaluation of a user model, that are complementary to the computation of its likelihood (or its perplexity):

1. At the limit, the correlation between the diagnostic and the prognostic shows how well the user model copes with the diversity of user behaviors;
2. The rate of convergence indicates how robust the metric is to new user sessions.

For the Deterministic Click model the utility is equal to the gain, while for the Probabilistic Click model, the utility is related to the gain by Eq. (8). We plot in Fig. 4 the prognostic DCG vs. the diagnostic DCG for the Deterministic Click (left) and Probabilistic (right) models. We have divided the prognostic DCG in 10 bins of equal range to ease visualization.[7] We also computed the Pearson correlation between the DCG and the diagnostic values of both models: The values are 15% for the Deterministic Click model and 34% for the Probabilistic model. This argues in favor of distinguishing the gain from the utility and once again argues in favor of the Probabilistic Click model.

6. Related works

Using user models in Information Retrieval has seen a much interest in the last few years, fostered first by needs of metrics related to the actual user experience in areas like XML retrieval [20], where more sophisticated user models started to appear.

[7] We had to withdraw the numerical values out of confidentiality concerns.

In Web search, most of the effort nowadays is to re-interpret standard metrics (AP and DCG). This, like in this paper, is essentially reduced to two sub-problems, namely defining the stopping criteria and a gain for each document the user inspects (according to the user model) [1].

Most of the effort has been focusing on the first problem (stopping criteria). Robertson [23] explicited the stopping criterion of AP and related it to the number of remaining relevant documents. Rank Biased Precision (RBP, defined first in [21]), is a metric that incorporates a "patience" parameter, i.e. a probability that the user continues, and a gain to each document that is seen by the user, thereby extending cumulated gains metrics. We have shown how the patience parameter can be understood as a parametric form of the DCG discounting factors of the model described in Section 3.

A difficult issue in designing more sophisticated user models is the problem of *setting* the parameters of the model. With the availability of user search click logs, it became easier to learn the parameters of the model directly from data. One of the first models to appear was proposed by Dupret et al. [11] in 2007 and Craswell et al. [5] in 2008. Many other work have been subsequently developed on the same basis. See for example Carterette et al. [3] who proposed to learn the *patience* parameter of RBP from a Bayesian perspective, or any of the work cited in the following paragraph.

Once a stopping criterion has been chosen, a utility can be computed. Chapelle et al. [4] propose to use the reciprocal rank and Yilmaz et al. [28] propose the EBU metric that depends both on user "patience" line RBP and on the last clicked document. Dupret and Piwowarski [13] propose a model where the stopping criterion depends on the whole set of clicked documents.

We discuss in more detail the work by Carterette [1], that introduces a framework for reasoning about probabilistic browsing models and IR measures, because it raises concerns about using Web search logs to validate metrics user models. In this work, he presents a series of parametric versions of DCG and compares them to other possible models as well as to an "empirical model" derived from clicks in a log. He puts up two arguments against the use of click log to learn the parameters. First, he argues that click log analysis does not allow to learn utility, which has already been addressed in the literature [4,28,13] by linking the stopping criterion to the amount of gain "collected" by the user. Second, he argues that maximizing likelihood might lead to a sub-optimal solution for some users, as a side effect maximizing for the whole set of users. In other words, the fact that experience might be much worse for some users is probably not compensated by making it only slightly better for a large number of other users. There are two complementary approaches to address this problem. First, as proposed in this paper, alternative criteria like comparing prognostic and diagnostic measurements allow to look at the discrepancy between the user model and the actual user behavior. Second, new measures could be derived from user models by looking not only at the expected value of the measure, but also at its variability: It is perfectly conceivable to analyze how the measure is distributed according to a user model rather than looking just at the mean (expected) value.

Other extensions of user models have tried to improve over diverse aspects. We discuss some of them below.

Handling graded relevance judgements. In order to solve this problem for recall-precision metrics, Piwowarski and Dupret [22], and then Robertson et al. [24], proposed to assume that users are defined by a relevance threshold under which they consider a document to be non-relevant, and by defining a simple probability distribution over these population of users.

Snippets. Incorporating snippet into the user model has been done by Turpin et al. [25], Yilmaz et al. [27]. They proposed to use the search result snippets into the user model, i.e. to incorporate the summary reading step of the search process, by making non-relevant documents on which the user would not have clicked because of the non-relevance of the snippet. We have mentioned in Section 3 how to include the snippets in the model and the same recipe applies to pAP. We didn't incorporate these into our numerical experiments as our objective in this work is not to propose new metrics but to understand better the two most popular, namely DCG and AP.

User models are not the only possibility when attempting to understand better and improve metrics. Kanoulas and Aslam [19] for example propose to chose the DCG gains and discount factors based on an analysis of variance approach whereby the relative variance due to the difference between ranking functions is maximized, the other variance components being due to topic (i.e. the choice of queries) variability and their interaction with the ranking functions.

7. Discussion

In this paper, we have illustrated the importance of defining user models for properly evaluating search engine performance, and we have compared two user models for DCG and one for a generalization of AP. We have seen how the DCG and AP metrics can be derived from a set of simple assumptions on how users interact with a ranking list.

Once a user model is chosen, it is possible to evaluate the model parameters as well as the associated metric. In particular, we have given estimates for the discounting factors of DCG. Even if it is not possible to estimate directly the gains, our analysis leads to an important practical finding: The popular PERFECT, EXCELLENT, GOOD, FAIR and BAD editorial labels used to evaluate Web search ranking, should not be associated with their "standard" gains (10, 7, 3, 0.5 and 0, respectively). This conclusion is based on our limited data set but it is likely to be true for Web search in general.

For DCG, we have seen that depending on the user model, we can interpret differently the gain (associated with a document) and discount factor (associated with a rank).

The Probabilistic Click model enables us to determine the discounting factors, but it doesn't help to determine the utilities $\mathrm{U}(\ell)$ or the gains G_ℓ. This is a consequence of a user model for which the relevance of the clicked documents have no influence on the user decision to stop the search. This is itself a consequence of the DCG definition: The discounting

factors D_r are the same no matter how relevant the documents at the other positions and therefore the contribution of the document at rank r to the final DCG value is independent of the other documents in the ranking.

The user models for DCG we presented in this work have in common the fact that the user decides before hand how many links in the ranking she will examine. As a consequence, the user decision to abandon the search is independent of her satisfaction. We have shown that on our data set, the pAP user model [13], which stopping criterion depends on how many relevant document the user has seen, and thus somehow on the satisfaction of the user, captures better the behavior of the user: The perplexity of the pAP model is significantly lower than that of DCG in our numerical experiments. Other data sets are likely to give different results, but it is interesting to have a criterion to determine which of two metrics is more adequate to a particular situation. The adequacy of a user model also depends on the search engine, because users might behave differently on different search engines.

Although we believe the Probabilistic Click Model is the best possible user model that can be associated with the DCG metric, it is always possible that other, more accurate user models exist. If this is the case, it should be easy to compare the models in term of their predictive ability and adjust the gains and discounting factors accordingly.

Finally, in this paper we have uncovered a new criterion to further classify which user model is better, in particular if their perplexity are similar. This criterion is based on the correlation between the prognostic (expectation of the metric given the ranking and the corresponding labels) and the diagnostic (expectation of the metric given the ranking, the labels *and the observed user behavior*). This correlation shows how well the diversity of user behavior is captured by the user model, as only in this case the diagnostic will converge towards the prognostic version of the metric. Hence, if there is an alternative user model for DCG, its superiority could be confirmed by verifying that its prognostic version is a better predictor of its diagnostic version than our Probabilistic Click Model.

Appendix A. Probabilistic Click model likelihood

The joint distribution describing the model is:

$$P(A=a, c_{1:R}, e_{1:R}; \ell_{1:R}) = P(A=a)\prod_{r=1}^{R} P(c_r|e_r;\ell_r)P(E_r|A)$$

where $E_{1:R}$ and A depend on each other deterministically. Hence, to compute the likelihood of a session, there is no need to sum over the states of $E_{1:R}$ because they are uniquely defined by the state of A.

Marginalizing over the possible states of A, we have

$$\begin{aligned}\mathcal{L}(\mathbf{s}) &= \sum_{a=1}^{R} P(A=a)\prod_{r=1}^{R} P(c_r|E_r;\ell_r)P(E_r|A=a)\\ &= \sum_{a=1}^{R} P(A=a)\left[\prod_{r=1}^{a} P\left(c_r|e_r^+;\ell_r\right)\times\prod_{r=a+1}^{R} P\left(c_r|e_r^-;\ell_r\right)\right]\end{aligned} \tag{A.1}$$

where we use the deterministic relationship between A and E.

From the assumptions of the user model, we know that a user cannot click a document at rank r if it has not been examined. This implies that the right-hand term is 0 if $r \geqslant b$ where b is the rank of the last click of the session $\mathbf{s}$. This in turn implies that Eq. (A.1) can be rewritten:

$$\mathcal{L}(\mathbf{s}) = \prod_{r=1}^{b} P\left(c_r|e_r^+;\ell_r\right)\sum_{a=b}^{R} P(A=a)\left[\prod_{r:a\geqslant r>b} P\left(c_r|e_r^+;\ell_r\right)\times\prod_{r=a+1}^{R} P\left(c_r^-|e_r^-;\ell_r\right)\right]$$

Then, we just need to observe that (a) at a rank $r \geqslant a+1 > b$ there is no click (by definition of b) and (b) if a rank is not examined, there is no click at this rank, i.e. that $P(c_r^-|e_r^-;\ell_r)=1$, which gives:

$$\mathcal{L}(\mathbf{s}) = \prod_{r=1}^{b} P\left(c_r|e_r^+;\ell_r\right)\sum_{a=b}^{R} P(A=a)\prod_{r/a\geqslant r>b} P\left(c_r|e_r^+;\ell_r\right)$$

Appendix B. Prognostic metric of the Probabilistic Click model

The expected utility can be expressed as

$$\mathbb{E}(\text{total utility}) = \mathbb{E}\left(\sum_{r=1}^{R}\mathrm{U}(\ell_r)c_r\right) = \sum_{r=1}^{R}\mathrm{U}(\ell_r)P\left(c_r^+\right)$$

where $c_r = 1$ if the document is clicked and 0 otherwise. We therefore need to compute the expectation of a click at an arbitrary position r. We can always write

$$\begin{aligned} P\left(c_r^+\right) &= P\left(c_r^+, A < r\right) + P\left(c_r^+, A \geqslant r\right) \\ &= 0 + P\left(c_r^+ | e_r^+; \ell_r\right) P(A \geqslant r) \end{aligned}$$

therefore

$$\mathbb{E}(\text{total utility}) = \sum_{r=1}^{R} \mathrm{U}(\ell_r) P\left(c_r^+ | e_r^+; \ell_r\right) P(A \geqslant r)$$

References

[1] B. Carterette, System effectiveness, user models, and user utility: A conceptual framework for investigation, in: SIGIR'11: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 2011, p. 903.

[2] B. Carterette, R. Jones, Evaluating search engines by modeling the relationship between relevance and clicks, Advances in Neural Information Processing Systems 20 (2008) 217–224.

[3] B. Carterette, E. Kanoulas, E. Yilmaz, Simulating simple user behavior for system effectiveness evaluation, in: The 20th ACM International Conference, ACM Press, New York, NY, USA, 2011, pp. 611–620.

[4] O. Chapelle, D. Metlzer, Y. Zhang, P. Grinspan, Expected reciprocal rank for graded relevance, CIKM'09: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM Request Permissions, 11 2009.

[5] N. Craswell, O. Zoeter, M. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: First ACM International Conference on Web Search and Data Mining WSDM 2008, 2008.

[6] F. Crestani, S. Marchand-Maillet, H.-H. Chen, E.N. Efthimiadis, J. Savoy (Eds.), Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010.

[7] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B 39 (1977) 1–38.

[8] S.C. Douglas, D. Beeferman, R. Rosenfeld, Evaluation metrics for language models, 1998.

[9] G. Dupret, User models to compare and evaluate Web IR metrics, in: Proceedings of SIGIR 2009 Workshop on the Future of IR Evaluation, http://staff.science.uva.nl/~kamps/ireval/papers/georges.pdf, 2009.

[10] G. Dupret, Discounted cumulative gain and user decision models, in: R. Grossi, F. Sebastiani, F. Silvestri (Eds.), SPIRE, in: Lecture Notes in Computer Science, vol. 7024, Springer-Verlag, 2011, pp. 2–13.

[11] G. Dupret, V. Murdock, B. Piwowarski, Web search engine evaluation using click-through data and a user model, in: Proceedings of the Workshop on Query Log Analysis (WWW), 2007.

[12] G. Dupret, B. Piwowarski, A user browsing model to predict search engine click data from past observations, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'08, ACM, New York, NY, USA, 2008, pp. 331–338.

[13] G. Dupret, B. Piwowarski, A user behavior model for average precision and its generalization to graded judgments, in: Crestani et al. [6], pp. 531–538.

[14] L. Granka, T. Joachims, G. Gay, Eye-tracking analysis of user behavior in www search, in: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2004, pp. 478–479.

[15] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, C. Faloutsos, Click chain model in Web search, in: Proceedings of the 18th International Conference on World Wide Web, WWW'09, ACM, New York, NY, USA, 2009, pp. 11–20.

[16] F. Guo, C. Liu, Y.M. Wang, Efficient multiple-click models in Web search, in: WSDM'09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2009, pp. 124–131.

[17] B. Hu, N.N. Liu, W. Chen, Learning from click model and latent factor model for relevance prediction challenge, in: Proceedings of the Second Workshop on Web Search Click Data (WSCD), 2012.

[18] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems (ACM TOIS) 20 (4) (2002) 222–246.

[19] E. Kanoulas, J. Aslam, Empirical justification of the gain and discount function for nDCG, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 611–620.

[20] G. Kazai, M. Lalmas, INEX 2005 evaluation measures, in: N. Fuhr, M. Lalmas, S. Malik, G. Kazai (Eds.), Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, 28–30 November, 2005, in: Lecture Notes in Computer Science, vol. 3977, Springer-Verlag, 2006, pp. 16–29.

[21] A. Moffat, J. Zobel, Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems 27 (1) (2008) 2:1–2:27.

[22] B. Piwowarski, G. Dupret, Evaluation in (XML) information retrieval: Expected precision-recall with user modelling (EPRUM), in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 8, 2006, ACM, 2006, pp. 260–267.

[23] S. Robertson, A new interpretation of average precision, in: SIGIR'08, ACM, New York, NY, USA, 2008, pp. 689–690.

[24] S.E. Robertson, E. Kanoulas, E. Yilmaz, Extending average precision to graded relevance judgments, in: Crestani et al. [6], pp. 603–610.

[25] A. Turpin, F. Scholer, K. Jarvelin, M. Wu, J.S. Culpepper, Including summaries in system evaluation, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'09, ACM, New York, NY, USA, 2009, pp. 508–515.

[26] E.M. Voorhees, D. Harman (Eds.), TREC: Experiment and Evaluation in Information Retrieval, MIT Press, 2005.

[27] E. Yilmaz, M. Shokouhi, N. Craswell, S. Robertson, Incorporating user behavior information in IR evaluation, in: Understanding the User SIGIR Workshop, 2009.

[28] E. Yilmaz, M. Shokouhi, N. Craswell, S. Robertson, Expected browsing utility for Web search evaluation, in: J. Huang, N. Koudas, G.J.F. Jones, X. Wu, K. Collins-Thompson, A. An (Eds.), CIKM, ACM, 2010, pp. 1561–1564.

[29] Y. Zhang, L. Park, A. Moffat, Click-based evidence for decaying weight distributions in search effectiveness metrics, Information Retrieval 13 (2010) 46–69, http://dx.doi.org/10.1007/s10791-009-9099-7.

[30] Y. Zhang, D. Wang, G. Wang, W. Chen, Z. Zhang, B. Hu, L. Zhang, Learning click models via probit bayesian inference, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10, ACM, New York, NY, USA, 2010, pp. 439–448.