A User Behavior Model for Average Precision and its Generalization to Graded Judgments

Georges Dupret
Yahoo! labs
gdupret@yahoo-inc.com

Benjamin Piwowarski
University of Glasgow
benjamin@bpiwowar.net

ABSTRACT

We explore a set of hypothesis on user behavior that are potentially at the origin of the (Mean) Average Precision (AP) metric. This allows us to propose a more realistic version of AP where users click non-deterministically on relevant documents and where the number of relevant documents in the collection needs not be known in advance. We then depart from the assumption that a document is either relevant or irrelevant and we use instead relevance judgment similar to editorial labels used for Discounted Cumulated Gain (DCG). We assume that clicked documents provide users with a certain level of "utility" and that a user ends a search when she gathered enough utility. Based on the query logs of a commercial search engine we show how to evaluate the utility associated with a label from the record of past user interactions with the search engine and we show how the two different user models can be evaluated based on their ability to predict accurately future clicks. Finally, based on these user models, we propose a measure that captures the relative quality of two rankings.

Categories and Subject Descriptors

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

General Terms

Theory, Experimentation

Keywords

Click-through Data, User Behavior, Search Engines, Statistical Model, Metrics

Introduction

An accurate method to quantify the quality of a document ranking is a fundamental requisite in the design of search engines. Ranking metrics intervene at different development stages: A *prognostic* metric is used to train a ranking function and to select the best one among a set of candidates. Once the function has been submitted to users, *diagnostic* metrics evaluate how users react to the changes brought by the new function.

Problem Description. Suppose reliable editors examined a set of documents returned in answer to a query and provided us, for each of them, with a label that describes its relevance on a five grades scale (in decreasing order of relevance): "PERFECT", "EXCELLENT", "GOOD", "FAIR" and "BAD" or P, E, G, F and B for short. As long as the user scans the ranking sequentially, i.e. from the top of the list to the bottom, one document at a time, it is clear that the best ranking is obtained by ordering the documents in decreasing order of their labels.

It is nevertheless not enough to know the ideal ranking. In a typical scenario, a new ranking function is designed to operate on a set of documents and query features. To compare this new function to a previous one or to evaluate it with respect to the optimal ranking, a random set of queries is chosen and the documents appearing in the rankings of both functions are manually labeled by editors. This gives rise to two sequences of labels for each query in the evaluation set. To compare these sequences, we have to work at two levels: **Individual Query Level:** Given the two sequences of ordered labels produced by two search engines in answer to a given query, which is more likely to satisfy user needs? **Ranking Function Level:** Supposing we know how to compare two rankings for a given query, how do we extend the results on individual queries to a set of queries? The second problem arises because when averaging over the results of several queries, it is not enough to know whether a query ranking is better than the other, it is also necessary to know by *how much*.

Contribution. We agree with Robertson [5] that "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". To illustrate the necessity of a user model, consider the case where a first ranking function produces the sequence $BBPBB$, while another function produces $FFFBB$. Provided users scan the list sequentially, if users stop their search after the second position in the ranking, then the second ranking is clearly better. This is not obviously true, and may even be false, if most of them scan at least three positions.

Resorting to user modeling also helps us break out the

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

"chicken and egg" problem [1] we face when comparing two different metrics: Deciding which metric is best calls for a third "meta" metric. Because various "meta" metrics are likely to co-exist, a meta metric for the meta metrics is necessary, etc. User models on the other hand can be compared based on their predictive ability. If one model predicts more accurately future user interactions with a search engine, then the metric derived from the first user model is arguably more reliable. This doesn't completely solve the problem though, as different metrics can be derived from a same user model.

This work first concentrates on describing a possible user model for an important and widely used metrics: Average Precision (AP, Section 1). This exercise will help us identify the implicit assumptions behind this metric and relate it quite naturally to other metrics found in the Literature. We will design the user model to be a fully generative statistical model based on explicit assumptions. This way it is possible to evaluate the model parameters based on past data (Section 1.5) and to evaluate the accuracy of the model on a test set. In Section 2, we propose an extension of AP to multi-graded relevance judgments and derive a new metric that compare two rankings based on the proportion of users that are better off with one ranking than with the other.

1. PROBABILISTIC AVERAGE PRECISION

The AP [6] metric can be associated to a particular set of hypothesis on the user behavior. This in turn defines a user model the parameters of which can be estimated from data. We will see that other metrics like pSkip [7], the Average Search Length and the Reciprocal Rank all share the same AP user model.

We first introduce some notations. Because we suppose that all documents are judged, we can understand a ranking as a sequence of labels $\ell_r, r = 1, \ldots, R$ where r indexes the position in the ranking. In the case of AP documents are either relevant (denoted by ℓ^+), or not relevant (ℓ^-).

We often use the notation $\ell_{1:R}$ to represent the whole ranking up to position R. We also introduce here the binary variable E_r, called the *examination*, that indicates whether a particular rank r is examined by the user. The subscript r is dropped when there is no ambiguity. By examining a position, we mean evaluating the snippet in order to decide whether the corresponding document is promising or not. 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 has been previously examined (There are no "accidental" clicks). On the other hand, if it is not clicked, we ignore if it was is examined or not.

In both our user models, we assume that a user is browsing sequentially a list, and that the user stops when satisfied. We define three different sets of boolean variables: S_r is true when the user is satisfied exactly at rank r, C_r is true when the user clicked the result at rank r, and E_r is true when the user examined the document at rank r. Note that the user is satisfied at one rank only, and hence boolean variable S_r can only be true for one rank r, if any: If S_r is true, then $S_{r'}$ is false for any rank r' different from r.

In order to keep notations compact, we use the following shorthands. First, for any random variable X, x^+ and x^- are equivalent to "X is *true*" and "X is *false*", respectively. We use lowercase x as a short-hand for $X = x$, and denote $\delta_{X=x}$ the indicator value which is 1 when the event $X = x$ is true, and 0 otherwise.

Another handy shorthand is used when we deal with a series of variables: A set of variables X_r for r between 1 and R is represented as $X_{1:R}$. Similarly $X_{a:b}$ is the set $X_a, X_{a+1}, \ldots, X_b$. This can be combined with the previous notations: $x^+_{a:b}$ is a short-hand for $x^+_a, x^+_{a+1}, \ldots, x^+_b$.

1.1 Average Precision

The AP metric is defined as the average of the precisions computed at the relevant document positions:

$$\texttt{AP} \quad = \quad \frac{1}{T} \sum_{r=1}^{\infty} \text{ precision at } r \times \text{ relevance at } r \qquad (1)$$

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 to a small number of terms.

1.2 User Model

To relate this measure to a user model, we first observe that the "precision at r" in Eq. 1 can be interpreted as a measure of how "easy" or "fast" r relevant documents are found by a user browsing the result list sequentially for exactly r relevant documents. If we further assume that $1/T$ users need exactly r relevant documents and that a user always clicks upon examining a relevant document, the expected precision coincide with AP, as discussed by Robertson [5]. In order to formalize these intuitions, we define the following user model:

User Model 1 (Probabilistic AP). [1]

1. *The user decides the number 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 that depends on the relevance of the document.*
4. *She ends her search as soon as she clicked on n relevant documents.*

Because different users need a different number of documents, n corresponds to discrete random variable that we denote by 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.

Several user models can be at the origin of the AP. For example, Moffat & Zobel [4] propose the next interpretation: "Every time a relevant document is encountered, the user pauses, asks "Over the documents I have seen so far, on average how satisfied am I" and writes a number on a piece of paper. Finally, when the user has examined every document in the collection –because this is the only way to be sure that all of the relevant ones have been seen– the user computes the average of the values they have written." This scenario stresses how unrealistic is the direct use the total number T of relevant documents as a component of an evaluation measure.

[1] pAP for short.

1.3 Prognostic Metric

Central to the metric is the evaluation of the rank at where the information need is met, which is described by $Pr(s_r^+)$.

According to the model assumption, the user browses the result list sequentially and abandons her search as soon as she meets her information need, so s_r^+ also implies that all documents are examined up to rank r and none is examined after rank r: $e_{1:r}^+$ and $e_{r+1:R}^-$. It also implies that the document at rank r was clicked, i.e. c_r^+.

A prognostic metric attempts to evaluate the quality of a ranking before it is presented to users. As a consequence, we need to evaluate $Pr(s_r^+)$ by marginalizing over all possible user interactions for all values of N ($Pr(s_r^+|n) = 0$. If $\ell_r = \ell^-$ then $Pr(s_r^+|n) = 0$). Otherwise:

$$\begin{aligned} Pr(s_r^+|n;\ell^+) &\stackrel{1}{=} \sum_{e_{1:R}}\sum_{c_{1:R}} \underbrace{Pr(s_r^+, e_{1:R}, c_{1:R}|n;\ell_{1:R})}_{\text{(a)}} \\ &\stackrel{2}{=} \sum_{c_{1:r-1}} Pr(s_r^+, e_{1:r}^+, e_{r+1:R}^-, c_{1:r-1}, c_r^+, c_{r+1:R}^-|n;\ell_{1:R}) \\ &\stackrel{3}{=} \sum_{c_{1:r-1}} \underbrace{Pr(e_{1:r}^+, e_{r+1:R}^-|s_r^+)}_{=1} \underbrace{Pr(c_r^+|e_r^+;\ell_r)}_{\text{(b)}} \underbrace{Pr(c_{r+1:R}^-|e_{r+1:R}^-)}_{=1} \\ &\quad \underbrace{Pr(c_{1:r-1}|e_{1:r-1}^+;\ell_{1:r-1})}_{\text{(c)}} \underbrace{Pr(s_r^+|c_{1:r-1}, c_r^+, n;\ell_{1:r-1})}_{\text{(d)}} \end{aligned} \quad (2)$$

The first equality holds by simple marginalization of the joint distribution over all the variables but S_r. Equality 2 holds because (a) is zero unless

1. $e_{1:r}^+$ and $e_{r+1:R}^-$ because s_r^+ entails that the user examined all positions up to r before ending the search at r,
2. c_r^+ because a user ends a search only if she clicks on a relevant document and
3. $c_{r+1:R}^-$ because there are no clicks on documents not examined.

Factor (d) is deterministic in Equality 3: It is zero unless

1. the document at position r is relevant
2. the user clicked at position r, an event that occurs with probability (b) if the document is relevant.
3. the user clicked on exactly $n-1$ relevant documents prior to position r, a condition that is realized with a probability that obeys to (c)

The model states that the probability to click does not depend on the rank, provided we know whether the user examined the position, and the label of the document at this rank. Hence, $Pr(c_r^+|e_r^+;\ell_r)$ is a constant independent of the rank; We denote it μ_+. We observe that if there are t_{r-1} relevant documents among the first $r-1$ positions, there are $\binom{n-1}{t_{r-1}}$ possible configurations[2], each of them having a probability

$$\mu_+^{n-1}(1-\mu_+)^{t_{r-1}-n+1}$$

[2]if $n-1 > t_{r-1}$ then there is no possible configuration, i.e. the value $\binom{n-1}{t_{r-1}}$ is defined as zero.

We have[3]:

$$\begin{aligned} Pr(s_r^+|n;\ell) &= \delta_{\ell_r^+}\mu_+ \binom{n-1}{t_{r-1}} \mu_+^{n-1}(1-\mu_+)^{t_{r-1}-n+1} \\ &= \delta_{\ell_r^+} \binom{n-1}{t_{r-1}} \mu_+^{n}(1-\mu_+)^{t_{r-1}-n+1} \end{aligned} \quad (3)$$

where the indicator $\delta_{\ell_r^+}$ is 1 if $\ell_r = \ell^+$ and 0 otherwise.

The first prognostic measure we define is the probabilistic interpretation of AP, that is the expected precision at the rank where the search is abandoned. In terms of our user model, this is:

$$\texttt{pAP}_{pro} = \mathbb{E}(\text{precision}) = \sum_n Pr(n) \sum_{r=1}^{R} \frac{n}{r} Pr(s_r^+|n)$$

Using Eq. 3, this can be expressed as:

$$\texttt{pAP}_{pro} = \sum_n Pr(n) \sum_{r=1}^{R} \delta_{\ell_r^+} \frac{n}{r} \binom{n-1}{t_{r-1}} \mu_+^{n}(1-\mu_+)^{t_{r-1}-n+1}$$

Finally, if we set $\mu_+ = 1$, $Pr(n) = T^{-1}$ for $n = 1, \ldots, T$ and $R = \infty$ we have:

$$\texttt{pAP}_{pro|\mu_+=1, Pr(n)=T^{-1}} = \sum_n \frac{1}{T} \sum_{r=1}^{\infty} \delta_{\ell_r^+} \delta_{t_r=n} \frac{n}{r}$$

This is the original AP as claimed above.

Various other prognostic metrics can be easily constructed based on the pAP user model once $Pr(s_r^+|N=n)$ defined in Eq. 3 is known. Maybe the most obvious are the *Expected Search Length* defined as

$$\texttt{ESL}_{pro} = \sum_r r Pr(s_r^+) = \sum_n Pr(n) \sum_r r Pr(s_r^+|n) \quad (4)$$

or the *Expected Reciprocal Rank*:

$$\texttt{ERR}_{pro} = \sum_r \frac{1}{r} Pr(s_r^+) = \sum_n Pr(n) \sum_r \frac{1}{r} Pr(s_r^+|n) \quad (5)$$

A definition closer to the original *Expected Search Length* from Cooper would consider the expected number of irrelevant documents before retrieving n relevant documents:

$$\sum_n Pr(n) \sum_r \frac{r-n}{r} Pr(s_r^+|n) \quad (6)$$

All these metrics are based on the knowledge of one central quantity, namely the probability $Pr(s_r^+|n)$ of the user being satisfied knowing she was looking for n relevant documents. This can be understood as different ways of weighting it for the rank. Unfortunately, although they are correlated these metrics do not necessarily lead to the same conclusion when used to compare two ranking functions.

1.4 Diagnostic Metric

Diagnostic metrics are meant to evaluate a ranking *after* it was presented to users and interactions have been collected. In the case of our user model, they are based on updating the probability $Pr(s_r;\ell_{1:R})$ with the click information, i.e. to estimate $Pr(s_r|c_{1:R};\ell_{1:R})$.

Suppose we observe a sequence of clicks and skips $c_{1:R}$ on a ranking defined by $\ell_{1:R}$. Given $c_{1:R}$, we know the number n_b of relevant documents clicked and the position b of

[3]This is the negative binomial distribution provided document at r is relevant.

the last click[4]. According to the user model, either the user was satisfied at rank b (s_b^+) or not satisfied at all ($s_{1:R}^-$). Hence, $Pr(s_{1:R}^-|c_{1:R})$ is simply $1 - Pr(s_b^+|c_{1:R})$ and all we need is to estimate $Pr(s_b^+|c_{1:R})$. If the clicked document is not relevant, this probability is 0. Otherwise, we know that if the user was satisfied at rank b, then she was looking for n_b relevant documents, and we have:

$$\begin{aligned} Pr(s_b^+|c_{1:R}) &= Pr(N = n_b|c_{1:R}) \\ &= \frac{Pr(c_{1:R}|n_b)Pr(n_b)}{Pr(c_{1:R}|n_b)Pr(n_b) + Pr(N > n_b)Pr(c_{1:R}|N > n_b)} \\ &= \frac{Pr(n_b)}{Pr(n_b) + Pr(N > n_b)\prod_{r=b+1}^{R} Pr(c_r^-|e_r^+;\ell_r)} \end{aligned}$$

Having defined the probabilities $Pr(s_r^+|c_{1:R}, \ell_{1:R})$, we can compute the diagnostic counterparts of Eqs. 4, 5 and 6 or any other suitable metric of interest. In particular, if we adopt the convention that the precision is null if the information need is not met (i.e. $s_{1:R}^-$), we can revise Eq. 4 and compute the expectation of precision *knowing the user clicks* $c_{1:R}$:

$$\texttt{pAP}_{dia} = \mathbb{E}(\text{precision}|c_{1:R}) = \delta_{\ell_b^+} Pr(s_b^+|c_{1:R})\frac{n_b}{b}$$

Moreover, if we suppose that all clicked documents are relevant and disregard the case where the user doesn't meet her information need, the "diagnostic" version of `pAP` coincide with `pSkip` [7] model with the `pSkip` metric being the empirical estimate of μ_+. Given the close relation of `pSkip` with the diagnostic versions of *Average Search Length* (`ASL`) and *Reciprocal Rank* (see [7]), we deduce that the `pAP` user model also generalizes the underlying user model of those metrics.

1.5 Parameters Estimation

In order to estimate the model parameters, we want to maximize the likelihood of the data. To define the latter, it is necessary to compute the likelihood of a session which is defined by the clicks $c_{1:R}$, i.e. to compute $\texttt{L} = Pr(c_{1:R}; \ell_{1:R})$.

Suppose that the last click of a session is at position b and that the user clicked on n_b relevant documents. In that case, we know that the user was either satisfied at rank b or continued his search beyond rank R. The likelihood of the first case is

$$\begin{aligned} &Pr(s_b^+, c_{1:R}; \ell_{1:R}) \\ &\qquad = Pr(s_b^+, c_{1:b}; \ell_{1:b}) = Pr(n_b)\prod_{r=1}^{b} Pr(c_r|e_r^+;\ell_r) \end{aligned}$$

while in the second case, the user is not satisfied by n_b relevant documents and:

$$Pr(s_b^-, c_{1:R}; \ell_{1:R}) = Pr(N > n_b)\prod_{r=1}^{R} Pr(c_r|e_r^+;\ell_r)$$

Because we don't observe S_b, we marginalize it to obtain

$$\texttt{L} = \quad Pr(n_b)\prod_{r=1}^{b} Pr(c_r|e_r^+;\ell_r) + Pr(N > n_b)\prod_{r=1}^{R} Pr(c_r|e_r^+;\ell_r)$$

A session without clicks is never satisfying for the user and its likelihood is obtained from the previous Equation by observing that $Pr(N = 0) = 0$ and $Pr(N > 0) = 1$.

[4]In the case there was no click, the model implies $s_{1:R}^-$.

To evaluate the probabilities $Pr(N = n)$, $\mu_+ = Pr(c^+|e^+;\ell^+)$ and $\mu_- = Pr(c^+|e^+;\ell^-)$ we multiply the likelihood of a set of the observed sessions and maximize the resulting product using standard techniques.

1.6 Numerical Experiments

We collected from the logs of the Yahoo! search engine a set of approximately 33,000 sessions with at least one click for which we have a PEGFB editorial judgment for each of the top 10 urls, together with a record of which urls have 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 $C_{1:10}$.

We divided the data in 10 random subsets and used each of these subsets (i.e. 10% of the original set) as the data we maximize the likelihood on. The data is labelled on the PEGFB scale and we need to decide how to adapt these 5 levels to the 2 levels –relevant and irrelevant– suitable for `pAP`. We explore successively all the possible mapping by considering first that only PERFECT documents are relevant (P row in Table 1), then that EXCELLENT and PERFECT documents are relevant (E case in Table 1), etc.

We observe that the estimates based on 10% of the data are fairly stable. If we consider all the documents with label "GOOD" or above as relevant, the probability of a click on a relevant document is 39% as opposed to only 19% on a irrelevant document. We also observe that 83% of users require 1 relevant document to satisfy their information need, while 12% need two and only 5% need more.

We would like to know which of the mappings from PEGFB labels to relevant or irrelevant is best aligned with the actual user behavior. The `pAP` model is a generative model and can be used to predict user behavior, i.e. which documents are clicked given a specific ranking; We can therefore compare the accuracy of these predictions on the test sets. We use the *perplexity* –a common measure of the "surprise" of a model when presented with a new observation. 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 D drawn from p. The perplexity of the model q is defined as

$$2^{-\sum_{i=1}^{D}\frac{1}{D}\log_2 q(x_i)} \tag{7}$$

Better approximations q of the unknown distribution p will tend to assign higher probabilities to the events observed in the test set. Thus, they have lower perplexity, i.e. they are "less surprised" by the test sample.

In the context of 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, the latter are held fixed at the values that maximize the likelihood on the training set.

All sessions in both the training and test sets contains $R = 10$ results so that by setting D to 10 times the number of sessions in Eq. 7, the perplexity is loosely[5] interpretable 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

[5]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.

Table 1: Median Click Probabilities and Required Number of Documents Distribution. The proportion of users requiring more than 4 documents is not significantly larger than zero.

	$Pr(c^+\|s^+;\ell)$		$Pr(N=n)$			
	μ_-	μ_+	$n=1$	$n=2$	$n=3$	$n=4$
B	-	0.42	0.89	0.11	0.00	0.00
F	0.17	0.44	0.88	0.12	0.00	0.00
G	0.19	0.39	0.83	0.12	0.03	0.02
E	0.14	0.33	0.90	0.07	0.02	0.01
P	0.12	0.68	0.99	0.01	0.00	0.00

to perfect predictions, while a perplexity of 2 corresponds to randomly predicting a click or a skip.

We have plotted the perplexity resulting from the 10 data splits for the 5 possible mappings in Figure 3. Experiments show that considering as relevant the document GOOD or better lead to the best model. To fix ideas, we also plotted the perplexity of a simple *CTR* (Click-Through Rate) model that predicts a click according to the CTR of the document label. For example, if 100 BAD labels appear in 50 sessions and are clicked 20 times, then the probability of a click on a BAD is estimated as $20/100 = 20\%$. This model doesn't take into account the document position in the ranking.

2. MULTI-GRADED MAP

The `pAP` model states that if a user needs n relevant documents, she will stop her search when she finds her n^{th} document and the documents beyond in the ranking have no importance to her. Although the assumption that a user stops her search as soon as her information need is met seems adequate, it is harder to believe that a pre-defined number of relevant documents will satisfy this need. It is also hard to believe that she actually knows this number. In the remaining of this work we propose a model where a certain amount of *utility* is associated to clicked documents, and a user stops her search when she gathered enough *utility* to meet her information need[6]. We relax the assumption that a document can either be relevant or not ($\ell \in \{\ell^+, \ell^-\}$) and allow multi-grade labels as for example `DCG` does. The user model is specified as follows:

User Model 2 (Satisfying Information Need). [7]

1. *The user examines the page results sequentially,*
2. *She clicks on a document she examines with a probability $Pr(c^+|e^+;\ell)$ that depends on the document label ℓ.*
3. *If she clicks on a document with label ℓ, she acquires the quantity $\mathtt{U}(\ell)$ of utility.*
4. *When she has gathered enough* utility *to satisfy her information need, she ends the search.*

We assume that utilities are additive: Each clicked document with label ℓ contributes an amount $\mathtt{U}(\ell)$ of utility to be added to the total utility the user already gathered. This is not completely realistic: If two documents provide the same content, the utility of consulting both should be the same as the utility of consulting one. We ignore this limitation in this work. As long as document relevances are judged independently from one another by editorial judges, there is no solution to this problem. Note that `AP` and `DCG` also suffers from this shortcoming.

We adopt the same notations as defined for the `pAP` user model (Section 1). The total utility associated with a set of clicks $c_{1:r}$ can be written $\sum_1^r \mathtt{U}_r \delta_{c_r}$ where δ_{c_r} is 1 if the user has clicked at rank r and 0 otherwise.

The larger the total utility the user acquires, the higher the probability that her information need is met. We capture the probabilistic relation between the total amount of utility – a continuous, positive variable – and the binary variable that states whether the user information need is met using the sigmoid function $\sigma(u) = (1 + \exp(-u_0 - u))^{-1}$ where u_0 is a suitable intercept. The effect of this function is to squash any value on the real axis to the interval $]0, 1[$ suitable for probabilities. With these assumptions, we are able to establish the relation between utility and information need:

$$Pr(s_r^+|s_{1:r-1}^-, c_{1:r}; \ell_{1:r}) \quad = \quad \delta_{c_r} \times \sigma(\sum_{s=1}^{r} \mathtt{U}_s c_s) \qquad (8)$$

where, as in Section 1, S_r is true if the user was satisfied at rank r. As the user clicks on more documents, the total utility increases, increasing the probability that the information need is met. Other parameterization are possible, but the logistic function presents some clear advantages: It is simple and it is monotonically increasing with its argument, the total amount of utility.

2.1 SIN Probabilities of Satisfaction

Prognostic Satisfaction. As for `AP`, we are interested in the rank where the user is satisfied $Pr(s_{1:R}; \ell_{1:R})$, which can be estimated by marginalizing the joint distribution of the model given by:

$$Pr(s_{1:R}, e_{1:R}, c_{1:R}; \ell_{1:R}) = \prod_{r=1}^{R} Pr(c_r|e_r;\ell_r) Pr(e_r|s_{1:r-1}) Pr(s_r|c_{1:r}, s_{1:r-1}; \ell_{1:r}) \qquad (9)$$

where the first component is estimated from the training data, the second is deterministic and the third is given by Eq. 8. This marginalization does not present any particular analytical difficulty, but we cannot expect the same kind of simplification as for `pAP`.

This process is illustrated in Figure 1 for the two first ranks: Assuming that the user always examines rank 1, she either clicks on the first document and follows the left branch –an event that happens with probability $Pr(c_1^+|e_1^+;\ell_1)$– or she skips it and follows the right path. If she chooses the first solution, she decides with probability $\sigma(\mathtt{U}_1)$ that the document at position 1 is sufficient to fulfill her information need and she ends her search. Otherwise she continues, an event that happens with probability $1 - \sigma(\mathtt{U}_1)$. Right before she reaches rank 2 she is in one of three states:

- She clicked on the document 1 and decided that her information need is satisfied. The search ends. (node 5 in Figure 1),
- She clicked on the document but decided her information need was not met (left most branch, node 3),

[6] This model is reminiscent of [2].
[7] `SIN` for short.

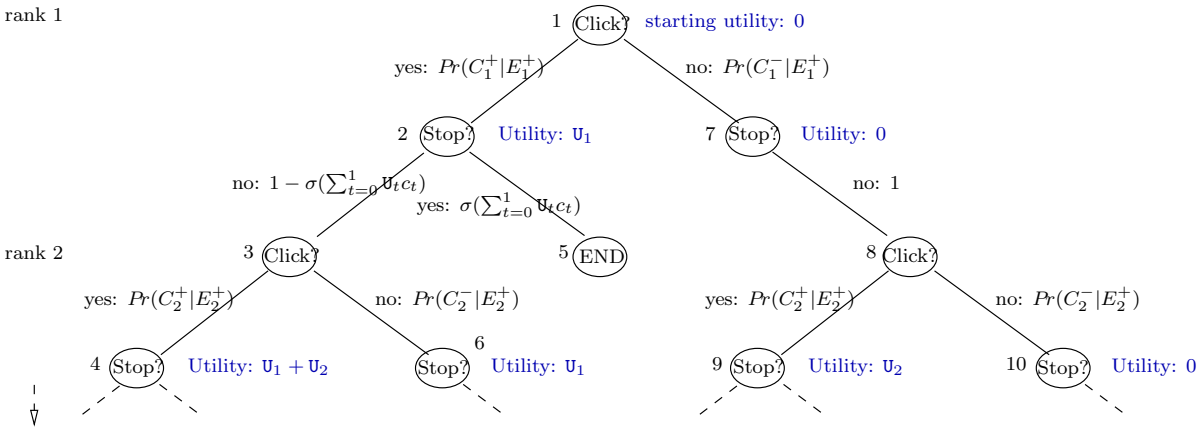

Figure 1: SIN decision process.

- She didn't click on the document and her information need is not met (right branch, node 8).

Each end node of rank 1 (nodes 3, 5 and 8) is reached with a probability equal to the product of the probabilities on the path from node 1 (reported in Figure 1 together with the *yes* / *no* decisions that determine the path). If the user didn't end at node 5, the process is repeated at node 3 and 8[8].

Diagnostic Satisfaction. In order to evaluate the diagnostic counterpart of the metrics, we need to estimate the probabilities of satisfaction *after* the user interactions have been observed. The development is similar to that of Section 1.5, where we distinguish two cases (the user was satisfied or not at the rank b of the last click). In the first case (s_b^+), we have

$$Pr(s_b^+, c_{1:R}; \ell_{1:R}) = Pr(s_b^+, s_{1:b-1}^-, e_{1:b}^+, e_{b+1:R}^-, c_{1:R}; \ell_{1:R})$$
$$= \prod_{r=1}^{b} Pr(c_r|e_r^+; \ell_r) Pr(s_r|c_{1:b}, s_{1:r-1}^-; \ell_{1:r}) \quad (10)$$

where we used Eq. 9 and the fact that (a) a user always examine the rank if she has not been satisfied before, i.e. $Pr(e_r^+|s_{1:r-1}^-) = 1$ for any rank less or equal than b, (b) a user does not examine any rank after being satisfied, i.e. $Pr(e_r^-|s_b^+) = 1$ for any rank after b and (c) she never clicks on non examined ranks, i.e. $Pr(c_r^-|e_r^-) = 1$ for any rank after b. Similarly, if the user is not satisfied, we have:

$$Pr(s_{1:R}^-, c_{1:R}; \ell_{1:R}) = Pr(s_{1:R}^-, e_{1:R}^+, c_{1:R}; \ell_{1:R})$$
$$= \prod_{r=1}^{R} Pr(c_r|e_r^+; \ell r) Pr(s_r^-|c_{1:r}, s_{1:r-1}^-; \ell_{1:r}) \quad (11)$$

and finally, combining Eqs. 10 and 11 :

$$Pr(s_b^+|c_{1:R}; \ell_{1:R}) = \frac{Pr(s_b^+, c_{1:R}; \ell_{1:R})}{Pr(s_b^-, c_{1:R}; \ell_{1:R}) + Pr(s_b^+, c_{1:R}; \ell_{1:R})}$$
$$= \frac{c_b \times \sigma(\sum_{1:R} \mathtt{U}_r c_r)}{\sigma(\sum_{1:R} \mathtt{U}_r c_r) + (1 - \sigma(\sum_{1:R} \mathtt{U}_r c_r)) \prod_{r=b+1}^{R} Pr(c_r^-|e_r^+; \ell r)}$$

where we used Eq. 8.

[8] A python script to compute $p(s_r^+; \ell_{1:r})$ is publicly available at: `http://sinmetric.sourceforge.net/`

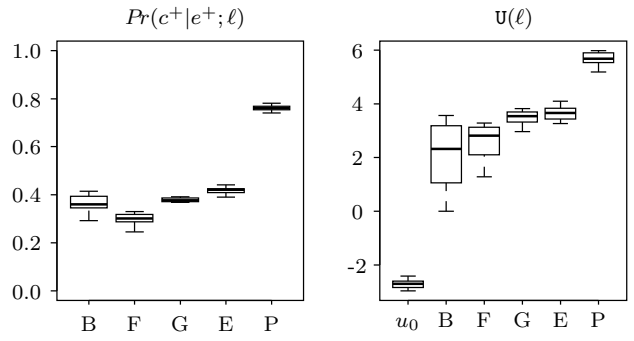

Figure 2: Left boxplot: Probability of click given the document label. Right boxplot: Utility of a document with the given label. The value of the intercept u_0 is also reported.

2.2 Model Estimation and Evaluation

The likelihood $\mathtt{L}$ of a session correspond to the probability $Pr(c_{1:R}; \ell_{1:R})$ of a sequence of clicks, which can be obtained by adding Eqs. 10 and 11:

$$Pr(c_{1:R}; \ell_{1:R}) = \prod_{r=1}^{b} Pr(c_r|e_r^+; \ell r) Pr(s_r|c_{1:r}, s_{1:r-1}^-; \ell_{1:r})$$
$$+ \prod_{r=1}^{R} Pr(c_r|e_r^+; \ell r) Pr(s_r^-|c_{1:r}, s_{1:r-1}^-; \ell_{1:r}) \quad (12)$$

As before, we maximize the likelihood of 10 subsets of our dataset. The results are reported in Figure 2 and Table 2. In our opinion these are very interesting results. First, click probabilities –with the exception of the label BAD– and utilities increase according to the label ordering, which corresponds to intuition but is not enforced by the model. The BAD documents have, when compared to FAIR documents, a higher probability of click but a lower utility. This is in agreement with what we expect of spam documents. Perfect documents –identified as the target of a navigational query– have a particularly high probability of being clicked and, if they are the first click, the user is predicted to stop with a probability $\sigma(\mathtt{U}(\text{PERFECT})) = 95\%$.

We computed the perplexity of the SIN model to compare

Table 2: Probability of click and utility according to the document label. The third column reports the probability of ending the search after clicking on one document with the corresponding label. The median of the intercept u_0 is -2.71.

Label ℓ	$Pr(c^+\|e^+;\ell)$	$\mathtt{U}(\ell)$	$\sigma(U(\ell))$
BAD	0.36	2.32	0.40
FAIR	0.30	2.81	0.52
GOOD	0.38	3.54	0.70
EXCELLENT	0.42	3.66	0.72
PERFECT	0.76	5.68	0.95

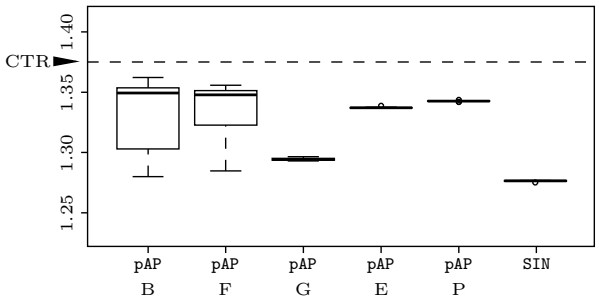

Figure 3: Perplexity of the different user models: Boxplots of 10 × cross-validation.

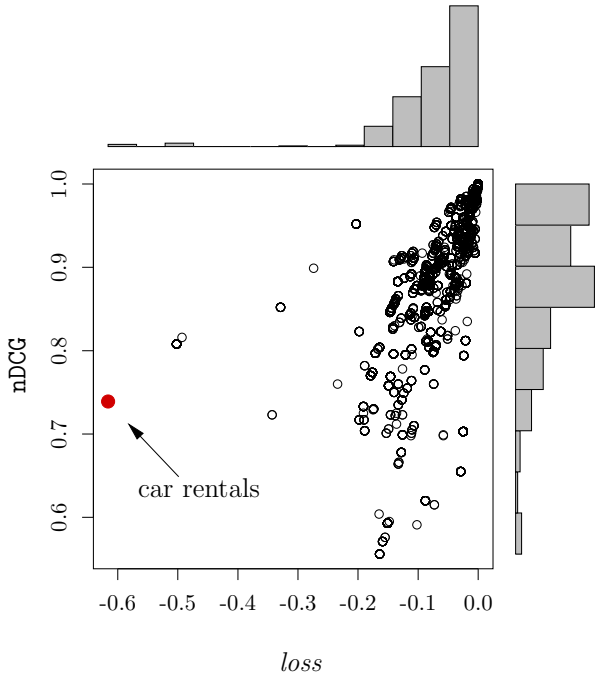

Figure 4: The *loss* vs. the nDCG for the ranking in the entire dataset. The histograms reflect the *loss* and nDCG values distribution (Top and Right, respectively).

it with the performance of the pAP user model. The results are shown in Figure 3 where we see that the SIN significantly outperforms pAP.

2.3 Loss & Benefit Metric

We can define the same set of prognostic and diagnostic metrics for the SIN as for pAP (Eqs. 4, 5 and 6), but this time the probability $Pr(s_r^+)$ is estimated based on the SIN user model.

Which of these metrics best reflects user satisfaction is not clear and might even be system dependent. These metrics can be understood as different ways of averaging the values of $Pr(s_r^+)$ over the positions r and are out of necessity to a certain extend arbitrary. The new approach we propose here doesn't avoid completely this caveat, but it attempts to remain as neutral as possible by avoiding the averaging over positions. As a basis for comparison between two rankings A and B, we propose to estimate the expected number of users who meet their information need earlier in one ranking than in the other. We denote S_A (resp. S_B) the set of satisfaction variables for ranking A (resp. B). The fact that a user meets her information need sooner in ranking A than in ranking B is denoted $A \succ B$ and happens with probability:

$$Pr(A \succ B) = \sum_{r=1}^{R} Pr(s^+_{A;r}, s^-_{B;1:r}) = \sum_{r=1}^{R} Pr(s^+_{A;r}) Pr(s^-_{B;1:r})$$

where we supposed that the user behavior on the two rankings were independent. The *benefit* $\mathcal{B}$ of ranking A –or the *loss* if negative– is defined as the proportion of users that are better off with ranking A than ranking B[9]:

$$\mathcal{B}(A,B) \quad = \quad Pr(A \succ B) - Pr(B \succ A)$$

[9]The same script of footnote 8 also computes the prognostic and the diagnostic *benefit*.

To compare two ranking functions, the *benefit* is averaged over queries.

We argue that this new metric is more "neutral" because no weighted average of $Pr(s_r^+)$ is computed for A or B. Instead the distributions of S^A and S^B are compared pointwise. The *benefit* also has the advantage of being straightforward to interpret.

2.4 Numerical Experiments

When comparing more than two systems, it is convenient to have an absolute value characterizing each of them separately. This is easily achieved by computing the *benefit* with respect to the ideal ranking obtained by ordering the documents according to their utilities. In Figure 4 we report the *benefit* with respect to the ideal ranking (actually a *loss*) as a function of the normalized DCG (nDCG) with logarithmic discounting factors. As expected, the larger the nDCG, the lower the *loss*. Although correlated, these two values are sufficiently different to lead to a different choice of ranking function.

The same Figure 4 distinguishes a set of ranking on the left for which the *loss* and nDCG lead to opposite conclusions: These rankings are very far from ideal in terms of the *loss* (values range from -.3 to -.5), while the nDCG is moderately below average with values in the .7 to .8 range. We isolate the ranking identified by a filled circle in Figure 4, extreme left to get more insight. It corresponds to the query "car rentals" with ranking GGEGGGPEGP. We report the different statistics associated with this ranking in Table 3. We first observe that the *loss* is stable from rank 2 because the vast majority of users (72.3 + 20.2 = 92.5% of users) presented with the ideal ranking meet their information need before

Table 3: Absolute gain for the query "car rentals". DCG discounts are logarithmic and scores are 10, 5, 3, 0.5, 0. The second column reports the document labels of the actual A and the ideal B rankings. The next two columns report the proportion of users meeting their information need at the different ranks for of rankings A and B, respectively.

rank	label	$Pr(S^A = r)$	$Pr(S^B = r)$	*benefit*	DCG	nDCG
1	G/P	0.265	0.723	-0.458	3.000	0.300
2	G/P	0.207	0.202	-0.549	4.893	0.300
3	E/E	0.176	0.025	-0.549	7.393	0.393
4	G/E	0.107	0.017	-0.550	8.685	0.414
5	G/G	0.076	0.010	-0.550	9.845	0.445
6	G/G	0.054	0.007	-0.550	10.914	0.471
7	P/G	0.085	0.005	-0.549	14.247	0.589
8	E/G	0.011	0.003	-0.549	15.825	0.630
9	G/G	0.006	0.002	-0.549	16.728	0.642
10	P/G	0.009	0.002	-0.549	19.618	0.729

rank 3 according to the SIN model. The *loss* is then essentially determined by the proportion of users who meet their information need on the actual ranking on the first two positions. In our data collection, a PERFECT document is the target page of a navigational query. The SIN model is consistent with this definition: It predicts with a high probability that the user will stop her search after seeing the target document. The nDCG on the other hand keeps increasing steadily up to rank 10 because the contribution of a given position to the final DCG value is independent of the documents presented at other positions.

Discussion

We have shown that a reasonable reconstruction of the user decision process can be deduced from the definition of AP. This is important because this help us question the implicit hypothesis behind this metric and propose the improvements at the origin of the pAP user model: We supposed first that users click on a document with a probability that depends on whether it is relevant or not, as opposed to AP where, at least implicitly, users always click on relevant documents. Like [4], we also reject the idea that the total number of relevant documents in the collection needs to be known to evaluate the system. Instead we suppose that the number of documents required by the user follows a distribution that can be estimated from past user interactions.

In the SIN model we further question the pAP hypothesis: Rather than supposing that users need a pre-defined number of relevant documents, we argued that they search as long as their information need is not satisfied. Making the assumption that documents with a higher level of relevance provide more "utility" to the user and contribute more to her satisfaction, we designed the SIN user model to predict user stops based on the total amount of utility she gathered. This hypothesis is more appealing intuitively and is able to handle naturally multi-graded relevance levels.

Unlike metrics, user models can be compared quantitatively because they have the ability to predict user interactions, i.e. which documents a user will click or skip when presented with a new ranking. By evaluating the prediction accuracy, we can determine which model is more adapted, i.e. represents better the user behavior, to a given search engine, a given market or a given set of users. In particular, we have shown that for our dataset the pAP model based on considering GOOD and better documents as relevant leads to the best prediction accuracy. We also showed that the SIN model outperforms the pAP model. This matched intuition because it is able to handle multi-graded levels of relevance.

Most metrics are based on the knowledge of a probability distribution on the rank at which the user meet her information need. This distribution can be evaluated prior to exposing the ranking to users by marginalizing over all the possible interactions with the result list. Metrics based on this prior distribution are qualified as *prognostic* metrics and can be used to train a ranking function or to chose among different candidates. The same probability distribution can be estimated after the new ranking function has been exposed to users and enough interactions have been recorded, giving rise the *diagnostic* counterpart of the metrics.

We have seen that a same user model can give rise to different metrics. Choosing one in particular is to a certain extent arbitrary and in this context it is important to make the weaker assumptions possible. This led us to propose that out of two rankings, the best is the one that leads the user to fulfill her information need at an earlier rank. Based on this definition, it is possible to estimate the *benefit* of a new ranking as the number of users it will favor.

The models we have proposed are still rather crude and many important aspects have been ignored. Both the pAP and the SIN models make the assumption that the user examines the ranking until she meets her information need. This is clearly unrealistic: Users do abandon search out of despair, reformulate their query, etc. The correction of this assumption is the topic of future work. Other aspects like document diversity, user diversity, query classes have also been ignored, etc. In this respect, the field of *Interactive Information Retrieval* [3] is certainly an important source of inspiration.

3. REFERENCES

[1] G. Dupret. User models to compare and evaluate web IR metrics. In *Proceedings of SIGIR 2009 Workshop on The Future of IR Evaluation*, 2009.

[2] G. Dupret and C. Liao. Estimating intrinsic document relevance from clicks. In *Proceedings of the 3rd WSDM conference*, 2010.

[3] D. Kelly. *Methods for Evaluating Interactive Information Retrieval Systems with Users*, volume 3 of *Foundations and Trends in Information Retrieval*. 2009.

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

[5] S. Robertson. A new interpretation of average precision. In *Proceedings of SIGIR'08*, pages 689–690, New York, NY, USA, 2008. ACM.

[6] E. M. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT press, 2005.

[7] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *Proceedings of the 15th ACM SIGKDD*, pages 1355–1364, New York, NY, USA, 2009. ACM.

More complete references for this work can be found in [1].