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 © 2012 Elsevier B.V. All rights reserved.
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, $r = 1, \dots, R$, 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: